Cisco Collaboration System 12.*x* Solution Reference Network Designs (SRND)

Last Updated: March 1, 2018

Cisco Systems, Inc.
www.cisco.com

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco website at www.cisco.com/go/offices.

First Published: February 7, 2017

THE SPECIFICATIONS AND INFORMATION REGARDING THE PRODUCTS IN THIS MANUAL ARE SUBJECT TO CHANGE WITHOUT NOTICE. ALL STATEMENTS, INFORMATION, AND RECOMMENDATIONS IN THIS MANUAL ARE BELIEVED TO BE ACCURATE BUT ARE PRESENTED WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. USERS MUST TAKE FULL RESPONSIBILITY FOR THEIR APPLICATION OF ANY PRODUCTS.

THE SOFTWARE LICENSE AND LIMITED WARRANTY FOR THE ACCOMPANYING PRODUCT ARE SET FORTH IN THE INFORMATION PACKET THAT SHIPPED WITH THE PRODUCT AND ARE INCORPORATED HEREIN BY THIS REFERENCE. IF YOU ARE UNABLE TO LOCATE THE SOFTWARE LICENSE OR LIMITED WARRANTY, CONTACT YOUR CISCO REPRESENTATIVE FOR A COPY.

The Cisco implementation of TCP header compression is an adaptation of a program developed by the University of California, Berkeley (UCB) as part of UCB's public domain version of the UNIX operating system. All rights reserved. Copyright © 1981, Regents of the University of California.

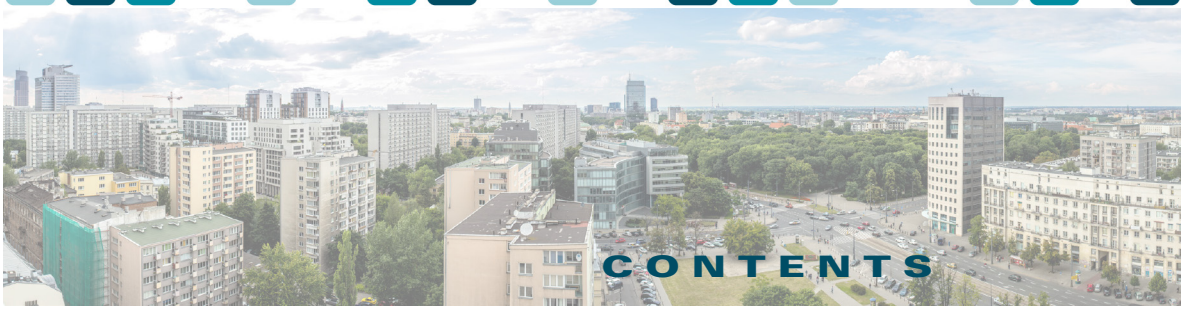
NOTWITHSTANDING ANY OTHER WARRANTY HEREIN, ALL DOCUMENT FILES AND SOFTWARE OF THESE SUPPLIERS ARE PROVIDED "AS IS" WITH ALL FAULTS. CISCO AND THE ABOVE-NAMED SUPPLIERS DISCLAIM ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING, WITHOUT LIMITATION, THOSE OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE.

IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THIS MANUAL, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1721R)

Any Internet Protocol (IP) addresses and phone numbers used in this document are not intended to be actual addresses and phone numbers. Any examples, command display output, network topology diagrams, and other figures included in the document are shown for illustrative purposes only. Any use of actual IP addresses or phone numbers in illustrative content is unintentional and coincidental.

Cisco Collaboration System 12.x SRND
© 2012-2018 Cisco Systems, Inc. All rights reserved.



Preface xxxvii

New or Changed Information for This Release xxxviii

Revision History xxxviii

Obtaining Documentation and Submitting a Service Request xxxviii

Cisco Product Security Overview xxxix

Conventions xxxix

CHAPTER 1 **Introduction** 1-1

Cisco End-to-End Collaboration Solutions 1-1

Collaboration Infrastructure 1-2

Collaboration Applications and Services 1-3

The Collaboration User Experience 1-4

About this Document 1-4

How this Document is Organized 1-5

Where to Find Additional Information 1-5

PART 1 **Collaboration System Components and Architecture**

CHAPTER 2 **Overview of Cisco Collaboration System Components and Architecture** 2-1

Architecture 2-3

High Availability 2-4

Capacity Planning 2-4

CHAPTER 3 **Network Infrastructure** 3-1

What's New in This Chapter 3-4

LAN Infrastructure 3-4

LAN Design for High Availability 3-4

Campus Access Layer 3-4

Routed Access Layer Designs 3-7

Campus Distribution Layer 3-9

Campus Core Layer 3-11

Power over Ethernet (PoE) 3-12

Energy Conservation for IP Phones 3-13

LAN Quality of Service (QoS) 3-14
Traffic Classification 3-16
Interface Queuing 3-18
Bandwidth Provisioning 3-19
Impairments to IP Communications if QoS is Not Employed 3-19
QoS Design Considerations for Virtual Unified Communications with Cisco UCS Servers 3-20
Congestion Scenario 3-20
QoS Implementation with Cisco UCS B-Series 3-20
QoS Design Considerations for Video 3-22
Network Services 3-23
Domain Name System (DNS) 3-23
Dynamic Host Configuration Protocol (DHCP) 3-24
Trivial File Transfer Protocol (TFTP) 3-28
Network Time Protocol (NTP) 3-33
WAN Infrastructure 3-33
WAN Design and Configuration 3-34
Deployment Considerations 3-34
Guaranteed Bandwidth 3-35
Dynamic Multipoint VPN (DMVPN) 3-35
Best-Effort Bandwidth 3-36
WAN Quality of Service (QoS) 3-37
WAN QoS Design Considerations 3-37
Considerations for Lower-Speed Links 3-45
Traffic Prioritization 3-47
Scavenger Class 3-48
Link Efficiency Techniques 3-48
Traffic Shaping 3-50
Bandwidth Provisioning 3-52
Provisioning for Bearer Traffic 3-53
Provisioning for Call Control Traffic 3-57
Wireless LAN Infrastructure 3-61
Architecture for Voice and Video over WLAN 3-62
Wireless Access Points 3-63
Wireless LAN Controllers 3-64
Authentication Database 3-64
Supporting Wired Network 3-64
Wireless Collaboration Endpoints 3-65
Wired Call Elements 3-65
Call Control 3-65
Media Termination 3-65

High Availability for Voice and Video over WLAN 3-66
Supporting Wired Network High Availability 3-66
WLAN High Availability 3-66
Call Processing High Availability 3-68
Capacity Planning for Voice and Video over WLAN 3-68
Design Considerations for Voice and Video over WLAN 3-68
Wireless AP Configuration and Design 3-72
Wireless LAN Controller Design Considerations 3-73
WLAN Quality of Service (QoS) 3-74
Traffic Classification 3-75
User Priority Mapping 3-75
Interface Queuing 3-76
Wireless Call Admission Control 3-77

CHAPTER 4 **Cisco Collaboration Security** 4-1
What's New in This Chapter 4-1
General Security 4-2
Security Policy 4-2
Security in Layers 4-3
Secure Infrastructure 4-4
Physical Security 4-4
IP Addressing 4-4
IPv6 Addressing 4-5
Access Security 4-5
Voice and Video VLANs 4-5
Switch Port 4-6
Port Security: MAC CAM Flooding 4-7
Port Security: Prevent Port Access 4-7
Port Security: Prevent Rogue Network Extensions 4-8
DHCP Snooping: Prevent Rogue DHCP Server Attacks 4-8
DHCP Snooping: Prevent DHCP Starvation Attacks 4-10
DHCP Snooping: Binding Information 4-10
Requirement for Dynamic ARP Inspection 4-11
802.1X Port-Based Authentication 4-12
Certificate Management 4-14
Brief PKI Overview 4-14
General Guidance on Certificates 4-15
RSA and ECDSA 4-16
CA-Signed Certificates Instead of Self-Signed Certificates 4-17

Multi-Server Certificates 4-18
Public versus Private CA 4-19
Encryption 4-19
TLS Overview 4-20
Cisco Unified CM Security 4-21
Unified CM Mixed Mode for Media and Signaling Encryption 4-22
Certificate Trust List (CTL) and Initial Trust List (ITL) 4-23
TFTP Configuration File Encryption 4-24
Survivable Remote Site Telephony (SRST) 4-25
Endpoint Security 4-25
PC Port on the Phone 4-26
PC Voice VLAN Access 4-26
Web Access Through the Phone 4-27
Settings Access 4-28
Cisco TelePresence Endpoint Hardening 4-28
Authentication and Encryption 4-29
VPN Client for IP Phones 4-30
Quality of Service 4-31
Access Control Lists 4-32
VLAN Access Control Lists 4-32
Router Access Control Lists 4-32
Firewalls 4-33
Routed ASA 4-35
Transparent ASA 4-36
Network Address Translation for Voice and Video 4-37
Data Center 4-37
Gateways, Trunks, and Media Resources 4-38
Putting Firewalls Around Gateways 4-39
Secure Audio and Video Conferencing 4-40
Unified CM Trunk Integration with Cisco Unified Border Element 4-40
Cisco Expressway in a DMZ 4-41
Applications Servers 4-41
Single Sign-On 4-42
SELinux on the Unified CM and Application Servers 4-42
General Server Guidelines 4-42
Deployment Examples 4-43
Lobby Phone Example 4-43
Firewall Deployment Example (Centralized Deployment) 4-44

Conclusion 4-44

CHAPTER 5 **Gateways 5-1**

What's New in This Chapter 5-1

Types of Cisco Gateways 5-2

Cisco TDM and Serial Gateways 5-2

Cisco Analog Gateways 5-2

Cisco Digital Trunk Gateways 5-3

Cisco TelePresence ISDN Link 5-3

TDM Gateway Selection 5-3

Gateway Protocols for Call Control 5-3

Core Feature Requirements 5-5

Gateways for Video Telephony 5-11

Dedicated Video Gateways 5-12

Integrated Video Gateways 5-13

Configuring Video Gateways in Unified CM 5-13

Call Signaling Timers 5-14

Bearer Capabilities of Cisco IOS Voice Gateways 5-14

IP Gateways 5-15

Cisco Unified Border Element 5-15

Cisco Expressway 5-16

Expressway-C and Expressway-E Deployment for Business-to-Business Communications 5-17

IP-Based Dialing for Business-to-Business Calls 5-20

High Availability for Expressway-C and Expressway-E 5-21

Security for Expressway-C and Expressway-E 5-23

Scaling the Expressway Solution 5-26

Considerations for Outbound Calls 5-31

Best Practices for Gateways 5-32

Tuning Gateway Gain Settings 5-32

Routing Inbound Calls from the PSTN 5-32

Gateway Digit Manipulation 5-33

Routing Outbound Calls to the PSTN 5-33

Video Gateway Call Bandwidth 5-34

Automated Alternate Routing (AAR) 5-34

Least-Cost Routing 5-36

Fax and Modem Support 5-37

CHAPTER 6 **Cisco Unified CM Trunks 6-1**

Unified CM Trunks Solution Architecture 6-2

A Comparison of SIP and H.323 Trunks 6-3
SIP Trunks Overview 6-5
Session Initiation Protocol (SIP) Operation 6-6
SIP Offer/Answer Model 6-6
SIP Delayed Offer 6-7
SIP Early Offer 6-7
Provisional Reliable Acknowledgement (PRACK) 6-8
Session Description Protocol (SDP) and Media Negotiation 6-9
Session Description Protocol (SDP) and Voice Calls 6-9
Session Description Protocol (SDP) and Video Calls 6-11
Video Desktop Sharing and Binary Floor Control Protocol (BFCP) 6-13
Far End Camera Control (FECC) 6-13
Unified CM SIP Trunk Features and Operation 6-14
Run on All Unified CM Nodes 6-14
SIP Trunks – Run on All Nodes and the Route Local Rule 6-14
Route Lists – Run on All Nodes and the Route Local Rule 6-15
Up to 16 SIP Trunk Destination IP Addresses 6-15
SIP Trunks Using DNS 6-17
SIP OPTIONS Ping 6-18
Unified CM SIP Trunks – Delayed Offer, Early Offer, and Best Effort Early Offer 6-18
Unified CM SIP Delayed Offer 6-18
Unified CM SIP Early Offer 6-19
Best Effort Early Offer [Early Offer support for voice and video calls Best Effort (no MTP inserted)] 6-22
MTP-Less Early Offer, Best Effort Early Offer, and SME Media Transparency 6-24
Media Termination Points 6-25
DTMF Transport over SIP Trunks 6-26
Codec Selection over SIP Trunks 6-28
Accept Audio Codec Preferences in Received Offer 6-30
Cisco Unified CM and Cisco Unified Border Element SIP Trunk Codec Preference 6-31
SIP Trunk Transport Protocols 6-32
Secure SIP Trunks 6-32
Media Encryption 6-32
Signaling Encryption 6-32
User Identity and SIP Trunks 6-34
Caller ID Presentation and Restriction 6-34
Called and Calling Party Number Normalization and SIP Trunks 6-35
Reasons for Using Only SIP Trunks in Cisco Collaboration Systems Deployments 6-36
Design and Configuration Recommendations for SIP Trunks 6-36

Unified CM Session Management Edition 6-38
When to Deploy Unified CM Session Management Edition 6-39
Differences Between Unified CM Session Management Edition and Standard Unified CM Clusters 6-40
Guidance on Centralizing Unified Communications Applications with Session Management Edition 6-42
Centralized Voice Mail – Unity Connection 6-43
Considerations for all QSIG Trunk Types 6-44
TelePresence Server and TelePresence Conductor 6-44
Expressway-C and Expressway-E 6-45
Summary of SIP Trunk Recommendations for Multi-Cluster SME Deployments 6-47
Minor Features of Unified CM SIP Trunks 6-50
SIP Trunk Message Normalization and Transparency 6-53
SIP Trunk Normalization 6-53
SIP Trunk Transparency 6-54
Pre-Loaded Unified CM Normalization and Transparency Scripts 6-55
IP PSTN and IP Trunks to Service Provider Networks 6-56
Cisco Unified Border Element 6-56
IP-PSTN Trunk Connection Models 6-57
IP PSTN Trunks and Emergency Services 6-59

CHAPTER 7 **Media Resources** 7-1
Media Resources Architecture 7-2
Media Resource Manager 7-2
Cisco IP Voice Media Streaming Application 7-3
Voice Termination 7-4
Medium and High Complexity Mode 7-4
Flex Mode 7-4
Transcoding 7-5
Audio Transcoding Resources 7-6
Video Interoperability 7-6
Media Termination Point (MTP) 7-7
Re-Packetization of a Stream 7-7
DTMF Conversion 7-7
DTMF Relay Between Endpoints 7-7
Calls over SIP Trunks 7-9
SIP Trunk MTP Requirements 7-11
DTMF Relay on SIP Gateways and Cisco Unified Border Element 7-12
H.323 Trunks and Gateways 7-12

H.323 Supplementary Services 7-12
H.323 Outbound Fast Connect 7-12
DTMF Conversion 7-13
DTMF Relay on H.323 Gateways and Cisco Unified Border Element 7-13
CTI Route Points 7-13
MTP Usage with a Conference Bridge 7-14
MTP Resources 7-14
Trusted Relay Point 7-15
Annunciator 7-15
Cisco RSVP Agent 7-17
Music on Hold 7-17
Unicast and Multicast MoH 7-17
MoH Selection Process 7-18
User and Network Hold 7-19
MoH Sources 7-21
Audio File 7-21
Fixed Source 7-21
Rebroadcast External Multicast Source 7-22
MoH Selection 7-23
MoH Call Flows 7-23
SCCP Call Flows 7-23
SIP Call Flows 7-26
Capacity Planning for Media Resources 7-30
Capacity Planning for Music on Hold 7-31
Co-resident and Standalone MoH 7-31
Server Platform Limits 7-31
Resource Provisioning 7-33
High Availability for Media Resources 7-34
Media Resource Groups and Lists 7-34
Redundancy and Failover Considerations for Cisco IOS-Based Media Resources 7-35
High Availability for Transcoders 7-36
High Availability for Music on Hold 7-36
Design Considerations for Media Resources 7-36
Deployment Models 7-36
Single-Site Deployments 7-36
Multisite Deployments with Centralized Call Processing 7-37
Multisite Deployments with Distributed Call Processing 7-38
Media Functions and Voice Quality 7-39
Music on Hold Design Considerations 7-39

Codec Selection 7-39
Multicast Addressing 7-39
Unified CM MoH Audio Sources 7-40
Unicast and Multicast in the Same Unified CM Cluster 7-40
Quality of Service (QoS) 7-41
Call Admission Control and MoH 7-41
Deployment Models for Music on Hold 7-43
Single-Site Campus (Relevant to All Deployments) 7-43
Centralized Multisite Deployments 7-43
Centralized PSTN Deployments 7-44
Multicast MoH from Branch Routers 7-44
Distributed Multisite Deployments 7-47
Clustering Over the WAN 7-47

CHAPTER 8 **Collaboration Endpoints 8-1**
What's New in This Chapter 8-2
Collaboration Endpoints Architecture 8-2
Cisco Unified Communications Manager (Unified CM) Call Control 8-4
Collaboration Endpoint Section 508 Conformance 8-5
Analog Endpoints 8-5
Standalone Analog Gateways 8-6
Analog Interface Module 8-6
Deployment Considerations for Analog Endpoints 8-6
Analog Connection Types 8-6
Paging Systems 8-7
Quality of Service 8-7
Desk Phones 8-8
Cisco Unified IP Phone 7900 Series 8-8
Cisco IP Phone 7800 Series 8-8
Cisco IP Phone 8800 Series 8-9
Cisco Unified SIP Phone 3900 Series 8-10
Cisco DX Series 8-10
Deployment Considerations for Cisco Desk Phones 8-11
Firmware Upgrades 8-11
Power Over Ethernet 8-12
Quality of Service 8-12
SRST and Enhanced SRST 8-13
Secure Remote Enterprise Attachment 8-13
Intelligent Proximity 8-13

Video Endpoints 8-14
Personal Video Endpoints 8-15
Cisco Jabber Desktop Video 8-15
Cisco IP Phone 8800 Series 8-15
Cisco DX Series 8-15
Cisco TelePresence System EX90 8-16
Multipurpose Video Endpoints 8-16
Cisco TelePresence System MX Series 8-16
Cisco TelePresence SX Series 8-17
Cisco Spark Room Series 8-17
Immersive Video Endpoints 8-18
Cisco TelePresence IX5000 Series 8-18
General Deployment Considerations for Video Endpoints 8-18
Quality of Service 8-18
Inter-VLAN Routing 8-19
SRST and Enhanced SRST 8-19
Secure Remote Enterprise Attachment 8-19
Intelligent Proximity 8-20
Video Interoperability 8-20
Software-Based Endpoints 8-22
Cisco IP Communicator 8-22
Cisco Jabber Desktop Clients 8-23
Cisco Jabber Desktop Client Architecture 8-23
Cisco Spark Desktop Clients 8-27
Cisco UC Integration™ for Microsoft Lync 8-27
Cisco UC Integration™ for Microsoft Lync Architecture 8-28
Deploying and Configuring Cisco UC Integration™ for Microsoft Lync 8-29
General Deployment Considerations for Software-Based Endpoints 8-29
Quality of Service 8-29
Inter-VLAN Routing 8-30
SRST and Enhanced SRST 8-30
Secure Remote Enterprise Attachment 8-30
Dial Plan 8-31
Contact Sources 8-32
Extend and Connect 8-32
OAuth with Refresh Login Flow 8-32
Wireless Endpoints 8-33
General Deployment Considerations for Wireless Endpoints 8-33
Network Radio Frequency Design and Site Survey 8-33
Security: Authentication and Encryption 8-34

Wireless Call Capacity 8-34
Bluetooth Support 8-35
Quality of Service 8-36
SRST and Enhanced SRST 8-36
Device Mobility 8-36
Mobile Endpoints 8-37
Cisco Jabber for Android and Apple iOS 8-37
Cisco Spark Mobile Clients 8-37
Cisco WebEx Meetings 8-38
Cisco AnyConnect Secure Mobility Client 8-38
Deployment Considerations for Mobile Endpoints and Clients 8-38
WLAN Design 8-38
Secure Remote Enterprise Attachment 8-38
Quality of Service 8-39
SRST and Enhanced SRST 8-40
Intelligent Proximity 8-40
Contact Sources 8-40
OAuth with Refresh Login Flow 8-41
Apple Push Notification Service (APNs) 8-41
Cisco Virtualization Experience Media Engine 8-42
Deployment Considerations for Cisco Virtualization Experience Media Engine 8-42
Quality of Service 8-42
SRST and Enhanced SRST 8-42
Third-Party IP Phones 8-43
High Availability for Collaboration Endpoints 8-43
Capacity Planning for Collaboration Endpoints 8-44
Design Considerations for Collaboration Endpoints 8-44

CHAPTER 9 **Call Processing** 9-1
What's New in This Chapter 9-2
Call Processing Architecture 9-2
Call Processing Virtualization 9-3
Call Processing Hardware 9-4
Unified CM Cluster Services 9-5
Cluster Server Nodes 9-6
Mixing Unified CM VM Configurations 9-8
Mixing Hardware Platforms and Business Edition Platforms 9-8
Intracluster Communications 9-9
Intracluster Security 9-11

General Clustering Guidelines 9-12
High Availability for Call Processing 9-13
Hardware Platform High Availability 9-13
Network Connectivity High Availability 9-13
Unified CM High Availability 9-14
Call Processing Redundancy 9-14
Call Processing Subscriber Redundancy 9-16
TFTP Redundancy 9-20
CTI Manager Redundancy 9-20
Virtual Machine Placement and Hardware Platform Redundancy 9-21
Cisco Business Edition High Availability 9-22
Capacity Planning for Call Processing 9-23
Unified CM Capacity Planning 9-23
Cisco Business Edition 6000M/H Capacity Planning 9-23
Cisco Business Edition 7000M/H and Cisco Unified CM Capacity Planning 9-24
Unified CM Capacity Planning Guidelines and Endpoint Limits 9-24
Megacluster 9-25
Cisco Business Edition 4000 Capacity Planning 9-26
Unified CME Capacity Planning 9-26
Design Considerations for Call Processing 9-26
Computer Telephony Integration (CTI) 9-28
CTI Architecture 9-29
CTI Applications and Clustering Over the WAN 9-31
Capacity Planning for CTI 9-32
High Availability for CTI 9-32
CTI Manager 9-32
Redundancy, Failover, and Load Balancing 9-32
Implementation 9-35
Integration of Multiple Call Processing Agents 9-36
Overview of Interoperability Between Unified CM and Unified CME 9-36
Call Types and Call Flows 9-36
Music on Hold 9-37
Instant and Permanent Hardware Conferencing 9-37
Unified CM and Unified CME Interoperability via SIP in a Multisite Deployment with Distributed Call Processing 9-38
Best Practices 9-38
Design Considerations 9-39

CHAPTER 10 **Collaboration Deployment Models** 10-1

What's New in This Chapter 10-1

Deploying Unified Communications and Collaboration 10-2

Deployment Model Architecture 10-4

Summary of Unified Communications Deployment Models 10-5

High Availability for Deployment Models 10-5

Capacity Planning for Deployment Models 10-6

Common Design Criteria 10-6

Site-Based Design Guidance 10-7

Centralized Services 10-8

Distributed Services 10-9

Inter-Networking of Services 10-9

Geographical Diversity of Unified Communications Services 10-9

Design Characteristics and Best Practices for Deployment Models 10-10

Campus Deployments 10-10

Best Practices for the Campus Model 10-12

Multisite Deployments with Centralized Call Processing 10-12

Best Practices for the Centralized Call Processing Model 10-16

Remote Site Survivability 10-16

Voice over the PSTN as a Variant of Centralized Call Processing 10-22

Multisite Deployments with Distributed Call Processing 10-23

Best Practices for the Distributed Call Processing Model 10-25

Leaf Unified Communications Systems for the Distributed Call Processing Model 10-25

Unified CM Session Management Edition 10-26

Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR) 10-32

Deployments for the Collaboration Edge 10-35

VPN Based Enterprise Access Deployments 10-35

VPN-less Enterprise Access 10-36

Business-to-Business Communications 10-37

IP PSTN Deployments 10-38

Design Considerations for Dual Call Control Deployments 10-40

Call Admission Control Considerations in Dual Call Control Deployments 10-41

Multisite Centralized Unified CM Deployments with Distributed Third-Party Call Control 10-41

Multisite Centralized Unified CM Deployments with Centralized Third-Party Call Control 10-42

Dial Plan Considerations in Dual Call Control Deployments 10-42

Clustering Over the IP WAN 10-43

WAN Considerations 10-44

Intra-Cluster Communications 10-45

Unified CM Publisher 10-45

Call Detail Records (CDR) and Call Management Records (CMR) 10-46
Delay Testing 10-46
Error Rate 10-47
Troubleshooting 10-47
Local Failover Deployment Model 10-47
Unified CM Provisioning for Local Failover 10-52
Gateways for Local Failover 10-53
Voicemail for Local Failover 10-53
Music on Hold and Media Resources for Local Failover 10-53
Remote Failover Deployment Model 10-54
Deploying Unified Communications on Virtualized Servers 10-55
Hypervisor 10-55
Server Hardware Options 10-56
Cisco Unified Computing System 10-56
Cisco UCS B-Series Blade Servers 10-56
Cisco UCS C-Series Rack-Mount Servers 10-58
Impact of Virtual Servers on Deployment Models 10-59
Call Routing and Dial Plan Distribution Using Call Control Discovery (CCD) for the Service Advertisement Framework (SAF) 10-59
Services that SAF Can Advertise with Call Control Discovery (CCD) 10-59
SAF CCD Deployment Considerations 10-60

CHAPTER 11 **Cisco Rich Media Conferencing** 11-1
What's New in This Chapter 11-3
Types of Conferences 11-3
Cisco Unified CM Audio Conferencing 11-4
Software Audio Conferencing 11-4
Hardware Audio Conferencing 11-5
Built-in Bridge 11-5
Cisco Conference Now 11-5
Cisco Meeting Server 11-7
Architecture 11-7
Role of Cisco Meeting Server 11-10
Role of Cisco TelePresence Management Suite (TMS) 11-12
Role of Cisco TelePresence Management Suite Extension for Microsoft Exchange (TMSXE) 11-12
Role of Cisco Meeting Management 11-12
Cisco Meeting Server Edge 11-12
Conference Call Flows 11-14

Instant Conferences 11-15
Permanent Conferences with Cisco Meeting Server Spaces 11-15
Scheduled Conferences 11-16
Security for Conferencing 11-18
High Availability for Conferencing 11-18
Cisco Unified CM High Availability 11-18
Cisco Meeting Server High Availability 11-19
Cisco TMS High Availability 11-21
Cisco Meeting Management High Availability 11-21
Scaling the Conferencing Solution 11-21
Considerations for Multiple Unified CM Clusters 11-22
Licensing 11-24
Capacity Planning 11-25
Design Considerations 11-26
Cisco WebEx Software as a Service 11-26
Architecture 11-26
Security 11-29
Scheduling 11-30
User Profile 11-30
High Availability 11-30
Cisco WebEx Cloud Connected Audio 11-31
Capacity Planning 11-33
Network Traffic Planning 11-33
Design Considerations 11-33
Cisco WebEx Meeting Center Video Conferencing 11-34
Architecture 11-34
Security 11-38
Audio Deployment Options 11-38
High Availability 11-39
Capacity Planning 11-39
Network Traffic Planning 11-40
Design Considerations 11-40
Cisco WebEx Meetings Server 11-41
Architecture 11-41
Cisco Unified CM Integration 11-44
Legacy PBX Integration 11-45
IPv6 Support 11-45
High Availability 11-46
Virtual IP Address 11-46

Multiple Data Center Design 11-46
Capacity Planning 11-47
Storage Planning 11-47
Network Traffic Planning 11-48
Design Consideration 11-48
Reference Document 11-49
Cisco Collaboration Meeting Rooms Hybrid 11-49
Architecture 11-50
Scheduling 11-52
Single Sign On 11-54
Security 11-54
Deployment Options 11-54
WebEx Audio Using SIP 11-55
WebEx Audio Using PSTN 11-55
Teleconferencing Service Provider Audio 11-57
High Availability 11-58
Capacity Planning 11-58
Network Traffic Planning 11-59
Design Considerations 11-59

PART 2 **Call Control and Routing**

CHAPTER 12 **Overview of Call Control and Routing** 12-1
Architecture 12-2
High Availability 12-3
Capacity Planning 12-3

CHAPTER 13 **Bandwidth Management** 13-1
What's New in This Chapter 13-1
Introduction 13-2
Collaboration Media 13-4
Fundamentals of Digital Video 13-4
Different Types of Video 13-4
H.264 Coding and Decoding Implications 13-4
Frame Types 13-5
Audio versus Video 13-6
Resolution 13-8
Network Load 13-9
Multicast 13-10

Transports 13-10
Buffering 13-11
Summary 13-11
"Smart" Media Techniques (Media Resilience and Rate Adaptation) 13-11
Encoder Pacing 13-11
Gradual Decoder Refresh (GDR) 13-12
Long Term Reference Frame (LTRF) 13-13
Forward Error Correction (FEC) 13-14
Rate Adaptation 13-15
Summary 13-15
QoS Architecture for Collaboration 13-16
Identification and Classification 13-17
QoS Trust and Enforcement 13-17
QoS for Cisco Jabber Clients 13-25
Utilizing the Operating System for QoS Trust, Classification, and Marking 13-29
Endpoint Identification and Classification Considerations and Recommendations 13-32
WAN Queuing and Scheduling 13-32
Dual Video Queue Approach 13-33
Single Video Queue Approach 13-34
Provisioning and Admission Control 13-37
Enhanced Locations Call Admission Control 13-39
Call Admission Control Architecture 13-40
Unified CM Enhanced Location Call Admission Control 13-40
Network Modeling with Locations, Links, and Weights 13-41
Location Bandwidth Manager 13-48
Enhanced Location CAC Design and Deployment Recommendations and Considerations 13-50
Intercluster Enhanced Location CAC 13-51
LBM Hub Replication Network 13-52
Common Locations (Shared Locations) and Links 13-53
Shadow Location 13-55
Location and Link Management Cluster 13-56
Intercluster Enhanced Location CAC Design and Deployment Recommendations and Considerations 13-58
Enhanced Location CAC for TelePresence Immersive Video 13-59
Video Call Traffic Class 13-59
Endpoint Classification 13-60
SIP Trunk Classification 13-60
Examples of Various Call Flows and Location and Link Bandwidth Pool Deductions 13-62
Video Bandwidth Utilization and Admission Control 13-66
Upgrade and Migration from Location CAC to Enhanced Location CAC 13-71

Extension Mobility Cross Cluster with Enhanced Location CAC 13-73
Design Considerations for Call Admission Control 13-73
Dual Data Center Design 13-74
MPLS Clouds 13-75
Call Admission Control Design Recommendations for Video Deployments 13-78
Enhanced Location CAC Design Considerations and Recommendations 13-80
Design Recommendations 13-80
Design Considerations 13-81
Design Recommendations for Unified CM Session Management Edition Deployments with Enhanced Location CAC 13-82
Recommendations and Design Considerations 13-82
Design Recommendations for Cisco Expressway Deployments with Enhanced Location CAC 13-85
Recommendations and Design Considerations 13-85
Design and Deployment Best Practices for Cisco Expressway VPN-less Access with Enhanced Location CAC 13-90
Bandwidth Management Design Examples 13-91
Example Enterprise #1 13-91
Identification and Classification 13-92
WAN Queuing and Scheduling 13-100
Provisioning and Admission Control 13-102
Example Enterprise #2 13-109
Identification and Classification 13-110
WAN Queuing and Scheduling 13-117
Provisioning and Admission Control 13-118

CHAPTER 14 **Dial Plan** 14-1
What's New in This Chapter 14-2
Dial Plan Fundamentals 14-3
Endpoint Addressing 14-3
Numeric Addresses (Numbers) 14-3
Alphanumeric Addresses 14-5
Dialing Habits 14-6
Dialing Domains 14-7
Classes of Service 14-8
Call Routing 14-8
Identification of Dialing Habit and Avoiding Overlaps 14-8
Forced On-Net Routing 14-10
Single Call Control Call Routing 14-11
Multiple Call Control Call Routing 14-11
Dial Plan Elements 14-13

Cisco Unified Communications Manager 14-13
Calling Party Transformations on IP Phones 14-14
Support for + Dialing on the Phones 14-15
User Input on SCCP Phones 14-15
User Input on Type-A SIP Phones 14-16
User Input on Type-B SIP Phones 14-18
SIP Dial Rules 14-20
Call Routing in Unified CM 14-22
Support for + Sign in Patterns 14-23
Directory URIs 14-23
Translation Patterns 14-24
External Routes in Unified CM 14-25
Pattern Urgency 14-36
Calling and Called Party Transformation Patterns 14-38
Incoming Calling Party Settings (per Gateway or Trunk) 14-40
Incoming Called Party Settings (per Gateway or Trunk) 14-41
Calling Privileges in Unified CM 14-41
Global Dial Plan Replication 14-47
Routing of SIP Requests in Unified CM 14-48
Cisco TelePresence Video Communication Server 14-53
Cisco VCS Addressing Schemes: SIP URI, H.323 ID, and E.164 Alias 14-53
Cisco VCS Addressing Zones 14-54
Cisco VCS Pattern Matching 14-54
Cisco VCS Routing Process 14-55
Recommended Design 14-56
Globalized Dial Plan Approach on Unified CM 14-56
Local Route Group 14-57
Support for + Dialing 14-57
Calling Party Number Transformations 14-58
Called Party Number Transformations 14-58
Incoming Calling Party Settings (per Gateway) 14-59
Logical Partitioning 14-60
Localized Call Ingress 14-61
Globalized Call Routing 14-62
Localized Call Egress 14-63
Call Routing in a Globalized Dial Plan 14-65
Benefits of the Design Approach 14-70
Dial Plan with Global Dial Plan Replication (GDPR) 14-72
Integrating Unified Communications Manager and TelePresence Video Communication Server 14-74

+E.164 Numbering Plan 14-75
Alias Normalization and Manipulation 14-75
Implementing Endpoint SIP URIs 14-78
Special Considerations 14-79
Automated Alternate Routing 14-79
Establish the PSTN Number of the Destination 14-80
Prefix the Required Access Codes 14-80
Voicemail Considerations 14-82
Select the Proper Dial Plan and Route 14-82
Device Mobility 14-83
Extension Mobility 14-84
Special Considerations for Cisco Unified Mobility 14-86
Remote Destination Profile 14-87
Remote Destination Profile's Rerouting Calling Search Space 14-87
Remote Destination Profile's Calling Search Space 14-87
Remote Destination Profile's Calling Party Transformation CSS and Transformation Patterns 14-88
Application Dial Rules 14-90
Time-of-Day Routing 14-91
Logical Partitioning 14-92
Logical Partitioning Device Types 14-93
Geolocation Creation 14-93
Geolocation Assignment 14-94
Geolocation Filter Creation 14-94
Geolocation Filter Assignment 14-94
Logical Partitioning Policy Configuration 14-94
Logical Partitioning Policy Application 14-95

CHAPTER 15 **Emergency Services** 15-1
911 Emergency Services Architecture 15-2
Public Safety Answering Point (PSAP) 15-2
Selective Router 15-2
Automatic Location Identifier Database 15-3
Service Provider ALI 15-3
Private Switch ALI 15-4
911 Network Service Provider 15-5
Interface Points into the Appropriate 911 Networks 15-6
Interface Type 15-7
Dynamic ANI (Trunk Connection) 15-7
Static ANI (Line Connection) 15-9

Cisco Emergency Responder 15-10
Device Location Discovery Methods in Cisco Emergency Responder 15-10
Switch Port Discovery 15-11
Access Point Association 15-11
IP Subnet 15-11
High Availability for Emergency Services 15-12
Capacity Planning for Cisco Emergency Responder Clustering 15-13
Design Considerations for 911 Emergency Services 15-13
Emergency Response Location Mapping 15-13
Emergency Location Identification Number Mapping 15-14
Dial Plan Considerations 15-16
Gateway Considerations 15-17
Gateway Placement 15-17
Gateway Blocking 15-17
Answer Supervision 15-18
Cisco Emergency Responder Design Considerations 15-19
Device Mobility Across Call Admission Control Locations 15-19
Default Emergency Response Location 15-19
Cisco Emergency Responder and Location Awareness for Wireless Clients 15-19
Cisco Emergency Responder and Extension Mobility 15-20
Cisco Emergency Responder and Video 15-20
Cisco Emergency Responder and Off-Premises Endpoints 15-21
Test Calls 15-21
PSAP Callback to Shared Directory Numbers 15-22
Cisco Emergency Responder Deployment Models 15-22
Single Cisco Emergency Responder Group 15-22
Multiple Cisco Emergency Responder Groups 15-24
Emergency Call Routing within a Cisco Emergency Responder Cluster 15-26
WAN Deployment of Cisco Emergency Responder 15-27
Emergency Call Routing Using Unified CM Native Emergency Call Routing 15-27
ALI Formats 15-29

CHAPTER 16 **Directory Integration and Identity Management** 16-1
What's New in This Chapter 16-2
What is Directory Integration? 16-3
Directory Access for Unified Communications Endpoints 16-4
Directory Access for Unified Communications Endpoints Using Cisco User Data Service (UDS) 16-6
Directory Integration with Unified CM 16-7

Cisco Unified Communications Directory Architecture 16-7
LDAP Synchronization 16-10
Synchronization Mechanism 16-14
Automatic Line Creation 16-17
Enterprise Group Support 16-19
Security Considerations 16-19
Design Considerations for LDAP Synchronization 16-19
Additional Considerations for Microsoft Active Directory 16-20
Unified CM Multi-Forest LDAP Synchronization 16-22
LDAP Authentication 16-22
Design Considerations for LDAP Authentication 16-25
Additional Considerations for Microsoft Active Directory 16-26
User Filtering for Directory Synchronization and Authentication 16-28
Optimizing Unified CM Database Synchronization 16-28
Using the LDAP Structure to Control Synchronization 16-29
LDAP Query 16-29
LDAP Query Filter Syntax and Server-Side Filtering 16-29
High Availability 16-31
Capacity Planning for Unified CM Database Synchronization 16-31
UDS Proxy for LDAP 16-32
Directory Integration for VCS Registered Endpoints 16-33
Identity Management Architecture Overview 16-33
Single Sign-On (SSO) 16-35
SAML Authentication 16-37
Authentication Mechanisms for Web-Based Applications 16-42
SSO for Cisco Jabber 16-43
Design Considerations for SSO 16-44
Authorization Framework 16-45
OAuth 2.0 16-45
OAuth Roles 16-46
General OAuth Flow 16-48
Authorization Grants 16-49
Tokens 16-51
Mobile and Remote Access (MRA) Authentication and Authorization 16-52
MRA Sign-On with Local Authentication 16-53
MRA Sign-On with SSO Authentication 16-54
Understanding OAuth Tokens 16-56
Access Tokens 16-56
Refresh Tokens 16-57

Token Signing and Encryption Keys 16-57
Scopes 16-57

PART 3 Collaboration Applications and Services

CHAPTER 17 Overview of Collaboration Applications and Services 17-1
Architecture 17-2
High Availability 17-3
Capacity Planning 17-4

CHAPTER 18 Cisco Unified CM Applications 18-1
What's New in This Chapter 18-2
IP Phone Services 18-2
IP Phone Services Architecture 18-2
High Availability for IP Phone Services 18-5
Capacity Planning for IP Phone Services 18-6
Design Considerations for IP Phone Services 18-7
Extension Mobility 18-7
Unified CM Services for Extension Mobility 18-8
Extension Mobility Architecture 18-8
Extension Mobility Cross Cluster (EMCC) 18-9
Call Processing 18-10
Media Resources 18-13
Extension Mobility Security 18-13
Support for Phones in Secure Mode 18-14
High Availability for Extension Mobility 18-15
Capacity Planning for Extension Mobility 18-17
Design Considerations for Extension Mobility 18-18
Design Considerations for Extension Mobility Cross Cluster (EMCC) 18-18
Unified CM Assistant 18-19
Unified CM Assistant Architecture 18-20
Unified CM Assistant Proxy Line Mode 18-20
Unified CM Assistant Share Lined Mode 18-21
Unified CM Assistant Architecture 18-22
High Availability for Unified CM Assistant 18-24
Service and Component Redundancy 18-24
Device and Reachability Redundancy 18-25
Capacity Planning for Unified CM Assistant 18-26
Design Considerations for Unified CM Assistant 18-28

Unified CM Assistant Extension Mobility Considerations 18-28
Unified CM Assistant Dial Plan Considerations 18-29
Unified CM Assistant Console 18-32
Unified CM Assistant Console Installation 18-32
Unified CM Assistant Desktop Console QoS 18-32
Unified CM Assistant Console Directory Window 18-33
Unified CM Assistant Phone Console QoS 18-33
WebDialer 18-34
WebDialer Architecture 18-34
WebDialer Servlet 18-34
Redirector Servlet 18-35
WebDialer Architecture 18-37
WebDialer URLs 18-38
High Availability for WebDialer 18-39
Service and Component Redundancy 18-40
Device and Reachability Redundancy 18-40
Capacity Planning for WebDialer 18-40
Design Considerations for WebDialer 18-41
Cisco Unified Attendant Consoles 18-42
Design Considerations for Cisco Unified Attendant Console Standard 18-43
Cisco Unified Attendant Console Advanced Architecture 18-43
High Availability for Cisco Unified Attendant Console Advanced 18-45
Design Considerations for Cisco Unified Attendant Console Advanced 18-45
Capacity Planning for Cisco Unified Attendant Consoles 18-47
Cisco Paging Server 18-47
Design Considerations for Cisco Paging Server 18-49

CHAPTER 19 **Cisco Voice Messaging** 19-1
Voice Messaging Portfolio 19-2
Messaging Deployment Models 19-3
Single-Site Messaging 19-4
Centralized Messaging 19-4
Distributed Messaging 19-4
Messaging and Unified CM Deployment Model Combinations 19-5
Cisco Unity Connection Messaging and Unified CM Deployment Models 19-6
Centralized Messaging and Centralized Call Processing 19-6
Cisco Unity Connection Survivable Remote Site Voicemail 19-8
Distributed Messaging with Centralized Call Processing 19-11
Combined Messaging Deployment Models 19-13

Centralized Messaging with Clustering Over the WAN 19-14
Distributed Messaging with Clustering Over the WAN 19-16
Messaging Redundancy 19-17
Cisco Unity Connection 19-17
Cisco Unity Connection Failover and Clustering Over the WAN 19-18
Cisco Unity Connection Redundancy and Clustering Over the WAN 19-19
Centralized Messaging with Distributed Unified CM Clusters 19-21
Cisco Unity Express Deployment Models 19-22
Overview of Cisco Unity Express 19-22
Deployment Models 19-22
Voicemail Networking 19-28
Cisco Unity Express Voicemail Networking 19-28
Interoperability Between Multiple Cisco Unity Connection Clusters or Networks 19-29
Cisco Unity Connection Virtualization 19-31
Best Practices for Voice Messaging 19-32
Best Practices for Deploying Cisco Unity Connection with Unified CM 19-32
Managing Bandwidth 19-32
Native Transcoding Operation 19-33
Cisco Unity Connection Operation 19-34
Integration with Cisco Unified CM 19-35
Integration with Cisco Unified CM Session Management Edition 19-36
IPv6 Support with Cisco Unity Connection 19-43
Single Inbox with Cisco Unity Connection 19-44
Best Practices for Deploying Cisco Unity Express 19-45
Voicemail Integration with Unified CM 19-45
Cisco Unity Express Codec and DTMF Support 19-46
JTAPI, SIP Trunk, and SIP Phone Support 19-46
Third-Party Voicemail Design 19-47

CHAPTER 20 **Collaboration Instant Messaging and Presence** 20-1
What's New in This Chapter 20-2
Presence 20-2
On-Premises Cisco IM and Presence Service Components 20-3
On-Premises Cisco IM and Presence Service User 20-3
Enhanced IM Addressing and IM Address Schemes 20-4
Single Sign-On (SSO) Solutions 20-4
IM and Presence Collaboration Clients 20-5
Multiple Device Messaging (MDM) and Logins 20-6
Jabber Desktop Client Modes 20-7

SAML Single Sign On 20-7
Cisco Unified CM User Data Service (UDS) 20-8
LDAP Directory 20-9
AD Groups and Enterprise Groups 20-9
AD Group Considerations for Groups and User Filters 20-10
WebEx Directory Integration 20-10
Common Deployment Models for Jabber Clients 20-10
On-Premises Deployment Model 20-11
Cloud-Based Deployment Model 20-12
Hybrid Cloud-Based and On-Premises Deployment Model 20-13
Client-Specific Design Considerations 20-14
Phone-Specific Presence and Busy Lamp Field 20-14
Unified CM Presence with SIP 20-14
Unified CM Speed Dial Presence 20-16
Unified CM Call History Presence 20-16
Unified CM Presence Policy 20-17
Unified CM Presence Guidelines 20-18
User Presence: Cisco IM and Presence Architecture 20-18
On-Premises Cisco IM and Presence Service Cluster 20-19
On-Premises Cisco IM and Presence Service High Availability 20-21
On-Premises Cisco IM and Presence Service Deployment Models 20-22
On-Premises Cisco IM and Presence Service Deployment Examples 20-23
On-Premises Cisco IM and Presence Service Performance 20-26
On-Premises Cisco IM and Presence Service Deployment 20-26
Single-Cluster Deployment 20-26
Intercluster Deployment 20-29
Clustering Over the WAN 20-29
Federated Deployment 20-36
On-Premises Cisco IM and Presence Service SAML SSO for Jabber 20-40
On-Premises Cisco IM and Presence Service Enterprise Instant Messaging 20-41
Deployment Considerations for Persistent Chat 20-42
Chat Room Limits for IM and Presence Service 20-43
Managed File Transfer 20-44
Managed File Transfer on IM and Presence Service 20-45
Managed File Transfer Capacities 20-45
On-Premises Cisco IM and Presence Service Message Archiving and Compliance 20-46
On-Premises Cisco IM and Presence Service Calendar Integration 20-51
Microsoft Outlook Calendar Integration 20-52
Multi-Language Calendar Support 20-53

Exchange Web Services Calendar Integration 20-53
On-Premises Cisco IM and Presence Service Mobility Integration 20-55
On-Premises Cisco IM and Presence Service Third-Party Open API 20-55
Design Considerations for On-Premises Cisco IM and Presence Service 20-58
Contact and Watcher List Recommendations 20-59
Mobile and Remote Access 20-61
Third-Party Presence Server Integration 20-62
Microsoft Communications Server for Remote Call Control (RCC) 20-62
In-the-Cloud Service and Architecture 20-64
Cisco WebEx Messenger 20-64
Deploying Cisco WebEx Messenger Service 20-64
Centralized Management 20-65
Single Sign On 20-66
Security 20-67
Firewall Domain White List 20-68
Logging Instant Messages 20-68
Capacity Planning for Cisco WebEx Messenger Service 20-68
High Availability for Cisco WebEx Messenger Service 20-69
Design Considerations for Cisco WebEx Messenger Service 20-69
Other Resources and Documentation 20-71

CHAPTER 21 **Mobile Collaboration** 21-1
What's New in This Chapter 21-3
Mobility Within the Enterprise 21-4
Campus Enterprise Mobility 21-4
Campus Enterprise Mobility Architecture 21-4
Types of Campus Mobility 21-5
Physical Wired Device Moves 21-5
Wireless Device Roaming 21-6
Extension Mobility (EM) 21-8
Campus Enterprise Mobility High Availability 21-9
Capacity Planning for Campus Enterprise Mobility 21-9
Design Considerations for Campus Enterprise Mobility 21-11
Multisite Enterprise Mobility 21-11
Multisite Enterprise Mobility Architecture 21-12
Types of Multisite Enterprise Mobility 21-13
Physical Wired Device Moves 21-13
Wireless Device Roaming 21-13
Extension Mobility (EM) 21-14

Device Mobility 21-14
Multisite Enterprise Mobility High Availability 21-24
Capacity Planning for Multisite Enterprise Mobility 21-25
Design Considerations for Multisite Enterprise Mobility 21-25
Remote Enterprise Mobility 21-26
Remote Enterprise Mobility Architecture 21-26
Types of Remote Enterprise Mobility 21-27
VPN Secure Remote Connectivity 21-28
Router-Based Remote VPN Connectivity 21-28
Client-Based Secure Remote Connectivity 21-28
Device Mobility and VPN Remote Enterprise Connectivity 21-29
VPN-Less Secure Remote Connectivity 21-30
Cisco Expressway 21-30
Remote Enterprise Mobility High Availability 21-32
Capacity Planning for Remote Enterprise Mobility 21-33
Design Considerations for Remote Enterprise Mobility 21-33
Cloud and Hybrid Services Mobility 21-34
Cloud and Hybrid Service Mobility Architecture 21-34
Types of Cloud Hybrid Service Integrations 21-36
Cisco WebEx Collaboration Cloud Hybrid Integrations 21-36
Cisco Spark Hybrid Services 21-36
Cloud and Hybrid Services Mobility High Availability 21-44
Capacity Planning for Cloud and Hybrid Services Mobility 21-45
Design Considerations for Cloud and Hybrid Services Mobility 21-45
Mobility Beyond the Enterprise 21-46
Cisco Unified Mobility 21-47
Single Number Reach 21-49
Single Number Reach Functionality 21-49
Single Number Reach Architecture 21-58
High Availability for Single Number Reach 21-58
Mobile Voice Access and Enterprise Feature Access 21-59
Mobile Voice Access IVR VoiceXML Gateway URL 21-60
Mobile Voice Access Functionality 21-60
Enterprise Feature Access with Two-Stage Dialing Functionality 21-63
Mobile Voice Access and Enterprise Feature Access Architecture 21-67
High Availability for Mobile Voice Access and Enterprise Feature Access 21-68
Designing Cisco Unified Mobility Deployments 21-68
Dial Plan Considerations for Cisco Unified Mobility 21-68
Guidelines and Restrictions for Unified Mobility 21-73

Capacity Planning for Cisco Unified Mobility 21-74
Design Considerations for Cisco Unified Mobility 21-74
Cisco Mobile Clients and Devices 21-76
Cisco Mobile Clients and Devices Architecture 21-77
Deployment Considerations for Cisco Mobile Clients and Devices 21-89
High Availability for Cisco Mobile Clients and Devices 21-109
Capacity Planning for Cisco Mobile Clients and Devices 21-110
Design Considerations for Cisco Mobile Clients and Devices 21-111

CHAPTER 22 **Cisco Unified Contact Center 22-1**
What's New in This Chapter 22-2
Cisco Contact Center Architecture 22-2
Cisco Unified CM Call Queuing 22-2
Cisco Unified Contact Center Enterprise 22-3
Cisco Unified Customer Voice Portal 22-4
Cisco Unified Contact Center Express 22-6
Cisco SocialMiner 22-6
Universal Queue for Third-Party Multichannel Applications 22-7
SocialMiner and Universal Queue 22-8
Unified CCE and Universal Queue 22-8
Finesse and Universal Queue 22-8
Administration and Management 22-8
Reporting 22-9
Multichannel Support 22-9
Recording and Silent Monitoring 22-9
Contact Sharing 22-10
Context Service 22-10
Cisco Virtualized Voice Browser 22-12
Contact Center Deployment Models 22-12
Single-Site Contact Center 22-12
Multisite Contact Center with Centralized Call Processing 22-12
Multisite Contact Center with Distributed Call Processing 22-14
Clustering Over the IP WAN 22-15
Design Considerations for Contact Center Deployments 22-17
High Availability for Contact Centers 22-17
Bandwidth, Latency, and QoS Considerations 22-18
Bandwidth Provisioning 22-18
Latency 22-19
QoS 22-19

Call Admission Control 22-19
Integration with Unified CM 22-20
Other Design Considerations for Contact Centers 22-20
Capacity Planning for Contact Centers 22-21
Video Customer Care 22-22
Cisco Remote Expert Solution 22-22
Network Management Tools 22-23

CHAPTER 23 **Call Recording and Monitoring** 23-1
What's New in This Chapter 23-1
Types of Monitoring and Recording Solutions 23-2
SPAN-Based Solutions 23-2
Unified CM Silent Monitoring 23-4
Unified CM Network-Based Recording 23-4
Unified CM Network-Based Recording with Built-in Bridge 23-6
Cisco Unified CM Network-Based Recording with a Gateway 23-7
Agent Desktop 23-10
Capacity Planning for Monitoring and Recording 23-10

PART 4 **Collaboration System Provisioning and Management**

CHAPTER 24 **Overview of Collaboration System Provisioning and Management** 24-1
Architecture 24-2
High Availability 24-3
Capacity Planning 24-3

CHAPTER 25 **Collaboration Solution Sizing Guidance** 25-1
What's New in This Chapter 25-2
Methodology for System Sizing 25-2
Performance Testing 25-2
System Modeling 25-3
Memory Usage Analysis 25-4
CPU Usage Analysis 25-4
Traffic Engineering 25-5
Definitions 25-5
Voice Traffic 25-6
Contact Center Traffic 25-7
Video Traffic 25-7

Conferencing and Collaboration Traffic 25-8
System Sizing Considerations 25-9
Network Design Factors 25-9
Other Sizing Factors 25-10
Sizing Tools Overview 25-10
Using the SME Sizing Tool 25-12
Using the VXI Sizing Tool 25-13
Using the Cisco Collaboration Sizing Tool 25-13
Cisco Unified Communications Manager 25-13
Virtual Nodes and Cluster Maximums 25-14
Deployment Options 25-14
Endpoints 25-16
Cisco Collaboration Clients and Applications 25-17
Call Traffic 25-22
Dial Plan 25-23
Applications and CTI 25-23
Media Resources 25-28
LDAP Directory Integration 25-31
Cisco Unified CM Megacluster Deployment 25-32
Cisco IM and Presence 25-33
Roster Management 25-34
Impact on Unified CM 25-35
Centralized IM and Presence 25-35
Emergency Services 25-36
Cisco Expressway 25-37
Gateways 25-38
Gateway Groups 25-38
PSTN Traffic 25-39
Gateway Sizing for Contact Center Traffic 25-39
Voice Activity Detection (VAD) 25-40
Codec 25-40
Performance Overload 25-40
Performance Tuning 25-41
Additional Information 25-42
Voice Messaging 25-42
Collaborative Conferencing 25-44
Sizing Guidelines for Audio Conferencing 25-44
Factors Affecting System Sizing 25-45
Sizing Guidelines for Video Conferencing 25-45

Impact on Unified CM 25-45
Cisco WebEx Meetings Server 25-45
Cisco Prime Collaboration Management Tools 25-48
Cisco Prime Collaboration Provisioning 25-48
Cisco Prime Collaboration Assurance 25-48
Cisco Prime Collaboration Analytics 25-49
Sizing for Standalone Products 25-49
Cisco Unified Communications Manager Express 25-49
Cisco Business Edition 25-49
Busy Hour Call Attempts (BHCA) for Cisco Business Edition 25-50
Cisco Unified Mobility for Cisco Business Edition 6000 25-52

CHAPTER 26 **Cisco Collaboration System Migration** 26-1
What's New in This Chapter 26-2
Coexistence or Migration of Solutions 26-2
Migration Prerequisites 26-3
Cisco Collaboration System Migration 26-3
Phased Migration 26-3
Parallel Cutover 26-3
Cisco Collaboration System Migration Examples 26-4
Summary of Cisco Collaboration System Migration 26-5
Centralized Deployment 26-5
Which Cisco Collaboration Service to Migrate First 26-6
Migrating Video Devices to Unified CM 26-6
Migrating Licenses to Cisco Collaboration System Release 12.*x* 26-7
License Migration with Cisco Global Licensing Operations (GLO) 26-7
Cisco Smart Software Manager 26-9
Using Cisco Prime Collaboration Deployment for Migration from Physical Servers to Virtual Machines 26-9
Cisco Prime Collaboration Deployment Migration Types 26-9
Cisco Prime Collaboration Deployment Migration Prerequisites 26-10
Simple Migration 26-10
Network Migration 26-10
Migrating Video Endpoints from Cisco VCS to Unified CM 26-11
Migrating from H.323 to SIP 26-11
Migrating Trunks from H.323 to SIP 26-11
Migrating Gateways from H.323 to SIP 26-12
Migrating Endpoints from SCCP to SIP 26-12

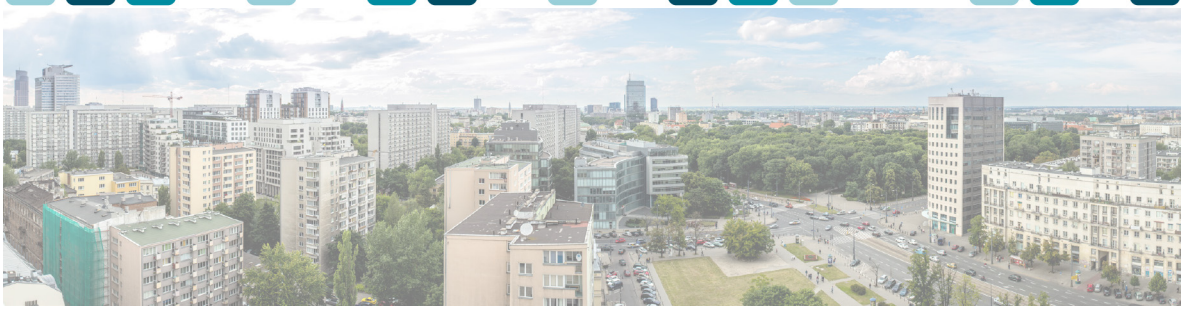
SIP URI Dialing and Directory Numbers 26-12
USB Support with Virtualized Unified CM 26-13
On-Premises Cisco IM and Presence Service Migration 26-14

CHAPTER 27 **Network Management** 27-1
What's New in This Chapter 27-2
Cisco Prime Collaboration 27-2
Failover and Redundancy 27-3
Cisco Prime Collaboration Server Performance 27-3
Network Infrastructure Requirements for Cisco Collaboration and Network Management Applications 27-4
Assurance 27-4
Assurance Design Considerations 27-7
Call Quality Monitoring (Service Experience) 27-8
Voice Quality Measurement 27-8
Unified CM Call Quality Monitoring 27-8
Cisco Network Analysis Module (NAM) 27-9
Comparison of Voice Quality Monitoring Methods 27-10
Trunk Utilization 27-10
Failover and Redundancy 27-10
Voice Monitoring Capabilities 27-10
Assurance Ports and Protocol 27-11
Bandwidth Requirements 27-12
Analytics 27-12
Analytics Server Performance 27-13
Provisioning 27-13
Provisioning Concepts 27-14
Best Practices 27-15
Prime Collaboration Design Considerations 27-16
Redundancy and Failover 27-17
Provisioning Ports and Protocol 27-17
Cisco TelePresence Management Suite (TMS) 27-18
Calendaring Options 27-18
Reporting 27-19
Management 27-19
Endpoint and Infrastructure Management 27-19
Provisioning 27-20
Phone books 27-20
Maintenance and Monitoring 27-21

Cisco Smart Software Licensing **27-21**
Deployment Scenarios **27-22**
Deployment Recommendations **27-23**
Redundancy **27-23**
Capacity Planning for Cisco Smart Software Manager **27-24**
Additional Tools **27-24**
Cisco Unified Analysis Manager **27-24**
Cisco Unified Reporting **27-25**
Integration with Cisco Collaboration Deployment Models **27-26**
Campus **27-26**
Multisite WAN with Centralized Call Processing **27-27**
Multisite WAN with Distributed Call Processing **27-28**
Clustering over the WAN **27-29**

GLOSSARY

INDEX



Preface

Revised: March 1, 2018

This document provides design considerations and guidelines for deploying Cisco Collaboration solutions, including Cisco Unified Communications Manager 12.*x*, Cisco TelePresence System, and other components of Cisco Collaboration System Release 12.*x*.

This document has evolved from a long line of Solution Reference Network Design (SRND) guides produced by Cisco over more than a decade. As Cisco's voice, video, and collaboration technologies have developed and grown over time, the SRND has been revised and updated to document those technology advancements. This latest version of the SRND includes Cisco's full spectrum of collaboration technologies such as Cisco Spark, TelePresence, WebEx, and support for a wide range of end-user devices. As Cisco continues to develop and enhance collaboration technologies, this SRND will continue to evolve and be updated to provide the latest guidelines, recommendations, and best practices for designing collaboration solutions.

This document should be used in conjunction with other documentation available at the following locations:

- For other Solution Reference Network Design (SRND) guides:

  https://www.cisco.com/go/srnd
- For information about Cisco Collaboration Preferred Architectures (PAs):

  https://www.cisco.com/go/pa
- For information about Cisco Collaboration Solutions:

  https://www.cisco.com/c/en/us/solutions/collaboration/index.html
- For information about Cisco Collaboration Systems Releases (CSRs):

  https://www.cisco.com/go/unified-techinfo
- For information about Cisco Unified Communications:

  https://www.cisco.com/c/en/us/products/unified-communications/index.html

  https://www.cisco.com/c/en/us/products/unified-communications/product-listing.html

  https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/tsd-products-support-series-home.html
- For information about Cisco Video Collaboration Solutions

  https://www.cisco.com/c/en/us/solutions/collaboration/video-collaboration/index.html

- For other Cisco design guides:

  https://www.cisco.com/go/designzone
- For all Cisco products and documentation:

  https://www.cisco.com

New or Changed Information for This Release

Note Unless stated otherwise, the information in this document applies to all Cisco Collaboration System 12.*x* releases.

Within each chapter of this guide, new and revised information is listed in a section titled *What's New in This Chapter*.

Although much of the content in this document is similar to previous releases of the *Cisco Collaboration SRND*, it has been reorganized and updated extensively to reflect more accurately the architecture of the current Cisco Collaboration System Release. Cisco recommends that you review this entire document, starting with the Introduction, page 1-1, to become familiar with the technology and the system architecture.

Revision History

This document may be updated at any time without notice. You can obtain the latest version of this document online at:

https://www.cisco.com/go/srnd

Visit the above website periodically and check for documentation updates by comparing the revision date of your copy with the revision date of the online document.

The following table lists the revision history for this document.

Revision Date	Comments
March 1, 2018	Initial version of this document for Cisco Collaboration System Release (CSR) 12.*x*.

Obtaining Documentation and Submitting a Service Request

For information on obtaining documentation, using the Cisco Bug Search Tool (BST), submitting a service request, and gathering additional information, see *What's New in Cisco Product Documentation*.

To receive new and revised Cisco technical content directly to your desktop, you can subscribe to the What's New in Cisco Product Documentation RSS feed. The RSS feeds are a free service.

Cisco Product Security Overview

This product contains cryptographic features and is subject to United States and local country laws governing import, export, transfer and use. Delivery of Cisco cryptographic products does not imply third-party authority to import, export, distribute, or use encryption. Importers, exporters, distributors and users are responsible for compliance with U.S. and local country laws. By using this product you agree to comply with applicable laws and regulations. If you are unable to comply with U.S. and local laws, return this product immediately.

Further information regarding U.S. export regulations may be found at:

https://www.access.gpo.gov/bis/ear/ear_data.html

Conventions

This document uses the following conventions:

Convention	Indication
bold font	Commands and keywords and user-entered text appear in **bold** font.
italic font	Document titles, new or emphasized terms, and arguments for which you supply values are in *italic* font.
[ ]	Elements in square brackets are optional.
{x | y | z }	Required alternative keywords are grouped in braces and separated by vertical bars.
[x | y | z]	Optional alternative keywords are grouped in brackets and separated by vertical bars.
string	A nonquoted set of characters. Do not use quotation marks around the string or the string will include the quotation marks.
`courier` font	Terminal sessions and information the system displays appear in `courier` font.
< >	Nonprinting characters such as passwords are in angle brackets.
[ ]	Default responses to system prompts are in square brackets.
!, #	An exclamation point (!) or a pound sign (#) at the beginning of a line of code indicates a comment line.

Note Means *reader take note*. Notes contain helpful suggestions or references to material not covered in the manual.

Tip Means *the following information will help you solve a problem*. The tips information might not be troubleshooting or even an action, but could be useful information, similar to a Timesaver.

Caution Means *reader be careful*. In this situation, you might perform an action that could result in equipment damage or loss of data.

Timesaver Means *the described action saves time*. You can save time by performing the action described in the paragraph.

Warning **IMPORTANT SAFETY INSTRUCTIONS**

This warning symbol means danger. You are in a situation that could cause bodily injury. Before you work on any equipment, be aware of the hazards involved with electrical circuitry and be familiar with standard practices for preventing accidents. Use the statement number provided at the end of each warning to locate its translation in the translated safety warnings that accompanied this device.

SAVE THESE INSTRUCTIONS

Warning **Statements using this symbol are provided for additional information and to comply with regulatory and customer requirements.**

Introduction

Revised: March 1, 2018

Collaboration means working together to achieve a common goal. Not very long ago, the best way for people to collaborate was for them to be in the same location at the same time so that they were in direct contact with each other. In today's globalized economy with decentralized business resources, outsourced services, and increasing costs for office facilities and travel, bringing people together in the same physical location is not the most efficient or effective way to collaborate. But with Cisco Collaboration Solutions, workers can now collaborate with each other anytime, anywhere, with substantial savings in time and expenses.

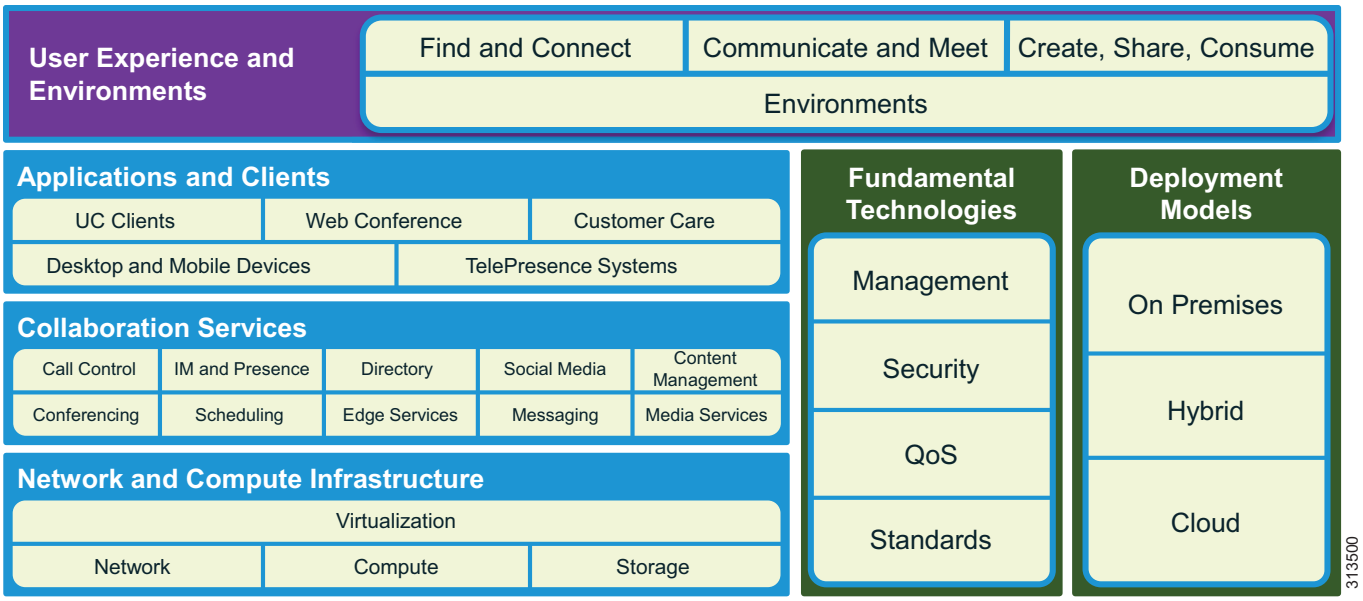
Cisco Collaboration Solutions support the full range of voice, video, and data communications, including the latest advances in mobile communications and social media. Cisco Collaboration Solutions also provide an extensive set of applications and services that can be deployed either on premises or in the cloud.

Cisco End-to-End Collaboration Solutions

Cisco Collaboration Technology comprises an array of products to build complete end-to-end collaboration solutions for virtually any size or type of enterprise. Cisco Collaboration Solutions consist of the following main elements, illustrated in conceptual form in Figure 1-1:

- Collaboration Infrastructure, page 1-2
- Collaboration Applications and Services, page 1-3
- The Collaboration User Experience, page 1-4

Figure 1-1 Cisco Collaboration Architecture (Conceptual View)



Collaboration Infrastructure

Cisco has long been recognized as the world leader in routing and switching technology. This technology forms the core of the network infrastructure for Cisco Collaboration Solutions. The Quality of Service (QoS) mechanisms available on Cisco switches and routers ensure that the voice, video, and data communications will be of the highest quality throughout the network. In addition, Cisco gateways provide a number of methods for connecting your enterprise's internal network to an external wide area network (WAN) as well as to the public switched telephone network (PSTN) and to legacy systems such as a PBX. And Cisco Hosted Collaboration Solution (HCS) enables Cisco partners to offer customers cloud-based, hosted collaboration services that are secure, flexible, low-cost, scalable, and always current with the latest technology.

Cisco Collaboration Systems Release 12.*x* is deployed using virtualization with the VMware vSphere ESXi Hypervisor. The Cisco Collaboration application nodes are deployed as virtual machines that can run as single or multiple application nodes on a server. These virtualized applications can provide collaboration services for small and medium businesses, and they can scale up to handle large global enterprises such as Cisco.

In most cases you will want your collaboration sessions to be secure. That is why Cisco has developed a number of security mechanism to protect each level of the collaboration path, from the network core to the end-user devices.

Once your collaboration solution is implemented, you will want to monitor and manage it. Cisco has developed a wide variety of tools, applications, and products to assist system administrators in provisioning, operating, monitoring and maintaining their collaboration solutions. With these tools the system administrator can monitor the operational status of network components, gather and analyze statistics about the system, and generate custom reports.

Collaboration Applications and Services

Cisco Collaboration Solutions incorporate a number of advanced applications and services, including:

- **Instant messaging (IM) and presence** — The Cisco IM and Presence Service enables Cisco Jabber, Cisco Unified Communications Manager applications, and third-party applications to increase user productivity by determining the most effective form of communication to help connect collaborating partners more efficiently.
- **Collaborative rich media conferencing** — Cisco WebEx incorporates audio, high-definition (HD) video, and real-time content sharing in a platform that provides easy setup and administration of meetings, interactive participation in meetings, and the ability to join meetings from any type of device such as an IP phone, a tablet device, or a desktop computer. For on-premises conferencing, Cisco TelePresence Server in combination with Cisco TelePresence Conductor enables ad hoc, scheduled, and permanent audio and video conferencing along with content sharing for TelePresence video endpoints, video-enabled desk phones, and software-based mobile and desktop clients.
- **Cisco Spark** — Cisco Spark desktop and mobile clients enable persistent cloud-based virtual team rooms that facilitate 1-to-1 and team collaboration. The Cisco Spark desktop client runs on Windows and Mac computers. The Cisco Spark mobile client runs on Android and Apple iOS devices. Cisco Spark allows users to access collaboration services from the Cisco Collaboration Cloud, including secure and encrypted persistent messaging, voice and video calls over IP, and file sharing, all within virtual one-on-one or group collaboration rooms.
- **Telepresence** — Cisco TelePresence technology brings people together in real-time without the expense and delay of travel. The Cisco TelePresence portfolio of products includes an array of high-definition (HD) video endpoints ranging from individual desktop units to large multi-screen immersive video systems for conference rooms. And Cisco TelePresence products are designed to interoperate with other Cisco collaboration products such as Cisco WebEx and Cisco IP Phones with video capability.
- **Voice messaging** — Cisco products provide several voice messaging options for large and small collaboration systems, as well as the ability to integrate with third-party voicemail systems using standard protocols.
- **Customer contact** — Cisco Unified Contact Center products provide intelligent contact routing, call treatment, and multichannel contact management for customer contact centers. Cisco Unified Customer Voice Portal can be installed as a standalone interactive voice recognition (IVR) system, or it can integrate with the contact center to deliver personalized self-service for customers. In addition, Cisco SocialMiner is a powerful tool for engaging with customers through social media.
- **Call recording and monitoring** — Cisco Collaboration Solutions can employ a variety of technologies to record and monitor audio and/or video conferences as well as customer conversations with contact center personnel. The call recording and monitoring technologies include solutions based on Cisco Unified Communications Manager, Cisco Agent Desktop, Cisco TelePresence Content Server, and Switched Port Analyzer (SPAN) technology.

The Collaboration User Experience

Collaboration is all about the user experience. When users have a good experience with collaboration technology, they will use that technology more often and will achieve better results with it. That translates into a bigger return on investment (ROI) for the enterprise that has adopted the collaboration technology. And that is why Cisco has focused on making its collaboration technology easy, convenient, and beneficial to use, with particular emphasis on the following enhancements to the user experience:

- **Wide variety of collaboration endpoints** — Cisco produces a compete line of endpoint devices ranging from basic voice-only phones, to phones with video and Internet capability, and to high-resolution telepresence and immersive video devices. Cisco Collaboration Technology also provides the ability to integrated third-party endpoint devices into the collaboration solution.
- **Cisco BYOD Smart Solution** — With the Cisco Bring Your Own Device (BYOD) Smart Solution, users can work from their favorite personal device, be it a smartphone, tablet, or PC. In addition to enhancing the work experience, the Cisco BYOD Smart Solution ensures greater network security and simplifies network management by providing a single policy for wired and Wi-Fi access across your organization.
- **Mobile collaboration** — Cisco mobile collaboration solutions provide mobile workers with persistent reachability and improved productivity as they move between, and work at, a variety of locations. Cisco mobility solutions include features and capabilities such as: Extension Mobility to enable users to log onto any phone in the system and have that phone assume the user's default phone settings; Cisco Jabber and Cisco Spark to provide core collaboration capabilities for voice, video, and instant messaging to users of third-party mobile devices such as smartphones and tablets; and Single Number Reach to provide a single enterprise phone number that rings simultaneously on an individual user's desk phone and mobile phone.
- **Applications and services** — As mentioned previously, Cisco has developed many advanced applications and services to enrich the collaboration experience for end users (see Collaboration Applications and Services, page 1-3). Whenever possible, Cisco strives to adhere to widely accepted industry standards in developing its collaboration technology so that you can easily integrate third-party applications and services into your collaboration solutions. In addition, the application programming interfaces available with many Cisco collaboration products enable you to develop your own custom applications.

About this Document

This document is a Solution Reference Network Design (SRND) guide for Cisco Collaboration Solutions. It presents system-level requirements, recommendations, guidelines, and best practices for designing a collaboration solution to fit your business needs.

This document has evolved from a long line of SRNDs produced by Cisco over more than a decade. As Cisco's voice, video, and data communications technologies have developed and grown over time, the SRND has been revised and updated to document those technology advancements. Early versions of the SRND focused exclusively on Cisco's Voice over IP (VoIP) technology. Subsequent versions documented Cisco Unified Communications and added information on new technologies for mobile voice communications, conferencing, instant messaging (IM), presence, and video telephony. This latest version of the SRND now includes Cisco's full spectrum of collaboration technologies such as Cisco Spark, TelePresence, and support for all types of end-user devices (Bring Your Own Device, or BYOD). As Cisco continues to develop and enhance collaboration technologies, this SRND will continue to evolve and be updated to provide the latest guidelines, recommendations, and best practices for designing collaboration solutions.

How this Document is Organized

This document is organized into four main parts:

- **Collaboration System Components and Architecture**

  The chapters in this part of the document describe the main components of Cisco Collaboration Technology and explain how those components work together to form a complete end-to-end collaboration solution. The main components include the network infrastructure, security, gateways, trunks, media resources, endpoints, call processing agents, deployment models, and rich media conferencing. For more information, see the Overview of Cisco Collaboration System Components and Architecture, page 2-1.

- **Call Control and Routing**

  The chapters in this part of the document explain how voice and video calls are established, routed, and managed in the collaboration system. The topics covered in this part include bandwidth management, dial plan, emergency services, and directory integration and identity management. For more information, see the Overview of Call Control and Routing, page 12-1.

- **Collaboration Applications and Services**

  The chapters in this part of the document describe the collaboration clients, applications, and services that can be incorporated into your collaboration solution. The topics covered in this part include Cisco Unified Communications Manager embedded applications, voice messaging, IM and presence, mobile collaboration, contact centers, and call recording. For more information, see the Overview of Collaboration Applications and Services, page 17-1.

- **Collaboration System Provisioning and Management**

  The chapters in this part of the document explain how to size the components of your collaboration solution, how to migrate to that solution, and how to manage it. The topics covered in this part include sizing considerations, migration options, and network management. For more information, see the Overview of Collaboration System Provisioning and Management, page 24-1.

Where to Find Additional Information

Because this document covers a wide spectrum of Cisco Collaboration products and possible solution designs, it cannot provide all the details of individual products, features, or configurations. For that type of detailed information, refer to the specific product documentation available at

https://www.cisco.com

This document provides general guidance on how to design your own collaboration solutions using Cisco Collaboration technology. Cisco has also developed, tested, and documented specific Preferred Architectures for collaboration, voice, and video deployments. The Preferred Architectures (PAs) provide prescriptive solution designs that are based on engineering best practices, and they are documented at

https://www.cisco.com/go/pa

PART 1

Collaboration System Components and Architecture

Contents of This Part

This part of the document contains the following chapters:

- Overview of Cisco Collaboration System Components and Architecture
- Network Infrastructure
- Cisco Collaboration Security
- Gateways
- Cisco Unified CM Trunks
- Media Resources
- Collaboration Endpoints
- Call Processing
- Collaboration Deployment Models
- Cisco Rich Media Conferencing

Overview of Cisco Collaboration System Components and Architecture

Revised: June 14, 2016

A solid network infrastructure is required to build a successful Unified Communications and Collaboration system in an enterprise environment. Other key aspects of the network architecture include selection of the proper hardware and software components, system security, and deployment models.

Unified Communications and Collaboration over an IP network places strict requirements on IP packet loss, packet delay, and delay variation (or jitter). Therefore, you need to enable most of the Quality of Service (QoS) mechanisms available on Cisco switches and routers throughout the network. For the same reasons, redundant devices and network links that provide quick convergence after network failures or topology changes are also important to ensure a highly available infrastructure. The following aspects are essential to the topic of Unified Communications and Collaboration networking and are specifically organized here in order of importance and relevance to one another:

- Network infrastructure — Ensures a redundant and resilient foundation with QoS enabled for Unified Communications and Collaboration applications.
- Voice security — Ensures a general security policy for Unified Communications and Collaboration applications, and a hardened and secure networking foundation for them to rely upon.
- Deployment models — Provide tested models for deploying Unified Communications and Collaboration call control and applications, as well as best practices and design guidelines to apply to Unified Communications and Collaboration deployments.

The chapters in this part of the SRND cover the networking subjects mentioned above. Each chapter provides an introduction to the subject matter, followed by discussions surrounding architecture, high availability, capacity planning, and design considerations. The chapters focus on design-related aspects rather than product-specific support and configuration information, which is covered in the related product documentation.

This part of the SRND includes the following chapters:

- Network Infrastructure, page 3-1

  This chapter describes the requirements of the network infrastructure needed to build a Cisco Unified Communications and Collaboration System in an enterprise environment. The sections in this chapter describe the network infrastructure features as they relate to LAN, WAN, and wireless LAN infrastructures. The chapter treats the areas of design, high availability, quality of service, and bandwidth provisioning as is pertinent to each infrastructure.

- Cisco Collaboration Security, page 4-1

  This chapter presents guidelines and recommendations for securing Unified Communications and Collaboration networks. The topics in this chapter range from general security, such as policy and securing the infrastructure, to endpoint security in VLANs, on switch ports, and with QoS. Other security aspects covered in this chapter include access control lists, securing gateways and media resources, firewalls, data center designs, securing application servers, and network virtualization.

- Gateways, page 5-1

  This chapter explores IP gateways, which are critical components of Unified Communications and Collaboration deployments because they provide the path for connecting to public networks. This chapter looks at gateway traffic types and patterns, protocols, capacity planning, and platform selection, as well as fax and modem support.

- Cisco Unified CM Trunks, page 6-1

  This chapter covers both intercluster and provider trunks, which provide the ability to route calls over IP and to leverage various Unified Communications and Collaboration features and functions. This chapter discusses H.323 and SIP trunks, codecs, and supplementary services over these trunks.

- Media Resources, page 7-1

  This chapter examines components classified as Unified Communications and Collaboration media resources. Digital signal processors (DSPs) and their deployment for call termination, conferencing and transcoding capabilities, and music on hold (MoH) are all discussed. Media termination points (MTPs), how they function, and design considerations with SIP and H.323 trunks are also covered. In addition, design considerations surrounding Trusted Relay Points, RSVP Agents, annunciator, MoH, and secure conferencing are included in the chapter.

- Collaboration Endpoints, page 8-1

  This chapter discusses the various types of Unified Communications and Collaboration endpoints available in the Cisco portfolio. Endpoints covered include software-based endpoints, wireless and hard-wired desk phones, video endpoints, and analog gateways and interface modules for analog connectivity based on time division multiplexing (TDM).

- Call Processing, page 9-1

  This chapter examines the various types of call processing applications and platforms that facilitate voice and video call routing. The chapter describes the call processing architecture, including platform options, clustering capabilities, and high availability considerations for call processing.

- Collaboration Deployment Models, page 10-1

  This chapter describes the deployment models for Cisco Unified Communications and Collaboration Systems as they relate to the various network infrastructures such as a single site or campus, multi-site environments, and data center solutions. This chapter covers these deployment models and the best practices and design considerations for each model, including many other subtopics pertinent to the model discussed.

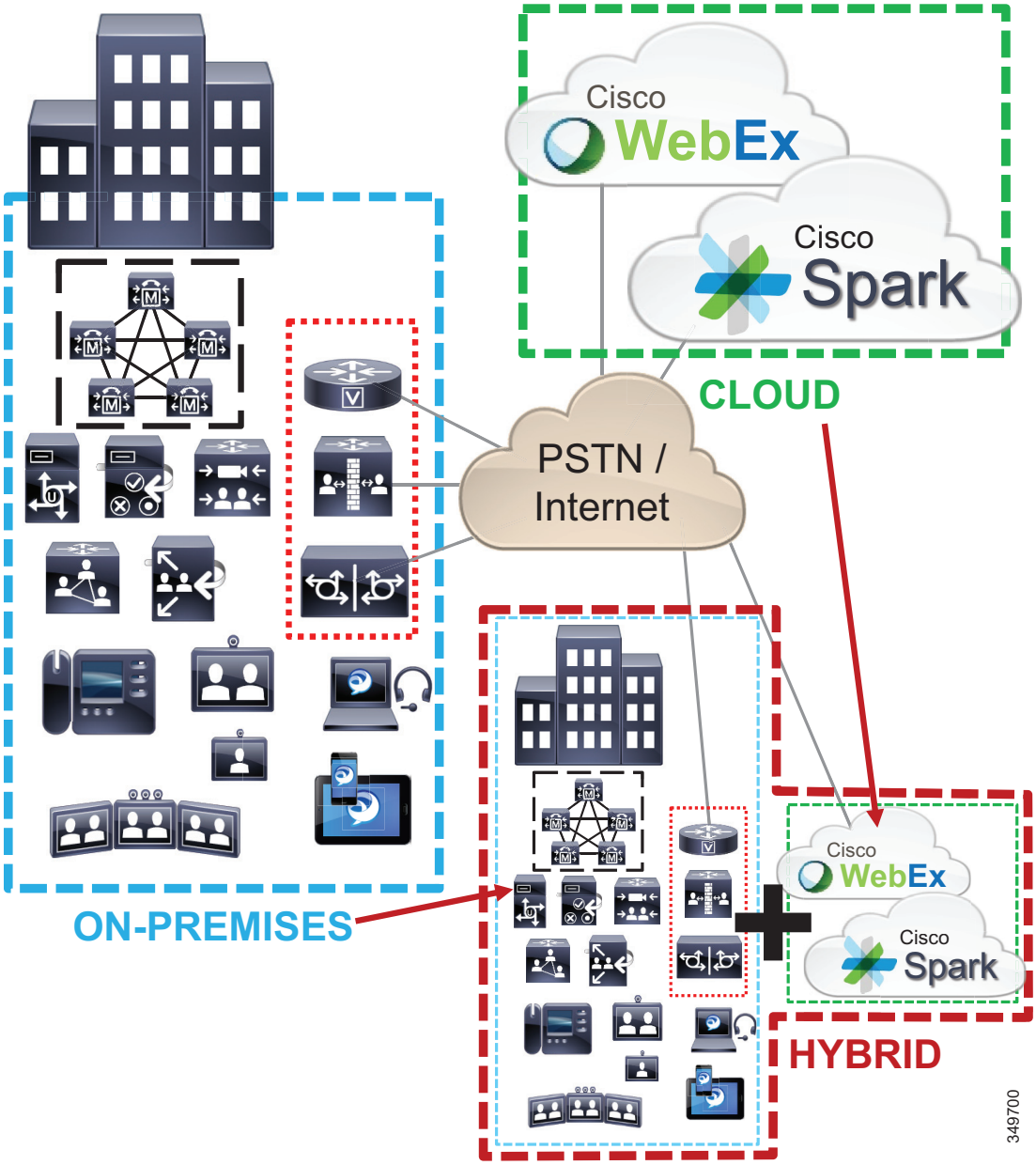
- Cisco Rich Media Conferencing, page 11-1

  This chapter explores rich media conferencing, which allows users of the Unified Communications and Collaboration system to schedule, manage, and attend audio, video, and/or web collaboration conferences. The chapter describes the different types of conferences as well as the software and hardware conferencing components, including the Cisco TelePresence Video Communication Server (VCS) and Multipoint Control Units (MCUs). The chapter also considers various aspects of rich media conferencing, such as deployment models, video capabilities, H.323 and SIP call control integrations, redundancy, and various solution recommendations and design best practices.

Architecture

The system architecture lays the foundation upon which all components of the Unified Communications and Collaboration System are deployed. Figure 2-1 illustrates, in a generalized way, how collaboration applications and services can be delivered solely on-premises, solely in the cloud, or in combination as a set of hybrid service deployments.

Figure 2-1 Enterprise Collaboration Deployments: On-Premises, Cloud, and Hybrid



All aspects of the Unified Communications and Collaboration System, including call routing, call control, applications and services, and operations and serviceability, rely heavily on proper design and deployment of the system architecture.

High Availability

Proper design of the network infrastructure requires building a robust and redundant network from the bottom up. By structuring the LAN as a layered model (access, distribution, and core layers) and developing the LAN infrastructure one step of the model at a time, you can build a highly available, fault tolerant, and redundant network. Proper WAN infrastructure design is also extremely important for normal operation on a converged network. Proper infrastructure design requires following basic configuration and design best-practices for deploying a WAN that is as highly available as possible and that provides guaranteed throughput. Furthermore, proper WAN infrastructure design requires deploying end-to-end QoS on all WAN links.

Wireless LAN infrastructure design becomes important when IP telephony is added to the wireless LAN (WLAN) portions of a converged network. With the addition of wireless Unified Communications and Collaboration endpoints, voice and video traffic has moved onto the WLAN and is now converged with the existing data traffic there. Just as with wired LAN and wired WAN infrastructures, the addition of voice and video in the WLAN requires following basic configuration and design best-practices for deploying a highly available network. In addition, proper WLAN infrastructure design requires understanding and deploying QoS on the wireless network to ensure end-to-end voice and video quality on the entire network.

After designing and implementing the network infrastructure properly, you can add network and application services successfully across the network, thus providing a highly available foundation upon which your Unified Communications and Collaboration services can run.

Capacity Planning

Scaling your network infrastructure to handle the Unified Communications and Collaboration applications and services that it must support requires providing adequate available bandwidth and the capability to handle the additional traffic load created by the applications.

For a complete discussion of system sizing, capacity planning, and deployment considerations related to sizing, refer to the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Network Infrastructure

Revised: March 1, 2018

This chapter describes the requirements of the network infrastructure needed to build a Cisco Unified Communications System in an enterprise environment. Figure 3-1 illustrates the roles of the various devices that form the network infrastructure, and Table 3-1 summarizes the features required to support each of these roles.

Unified Communications places strict requirements on IP packet loss, packet delay, and delay variation (or jitter). Therefore, it is important to enable most of the Quality of Service (QoS) mechanisms available on Cisco switches and routers throughout the network. For the same reasons, redundant devices and network links that provide quick convergence after network failures or topology changes are also important to ensure a highly available infrastructure

The following sections describe the network infrastructure features as they relate to:

- LAN Infrastructure, page 3-4
- WAN Infrastructure, page 3-33
- Wireless LAN Infrastructure, page 3-61

Figure 3-1 Typical Campus Network Infrastructure

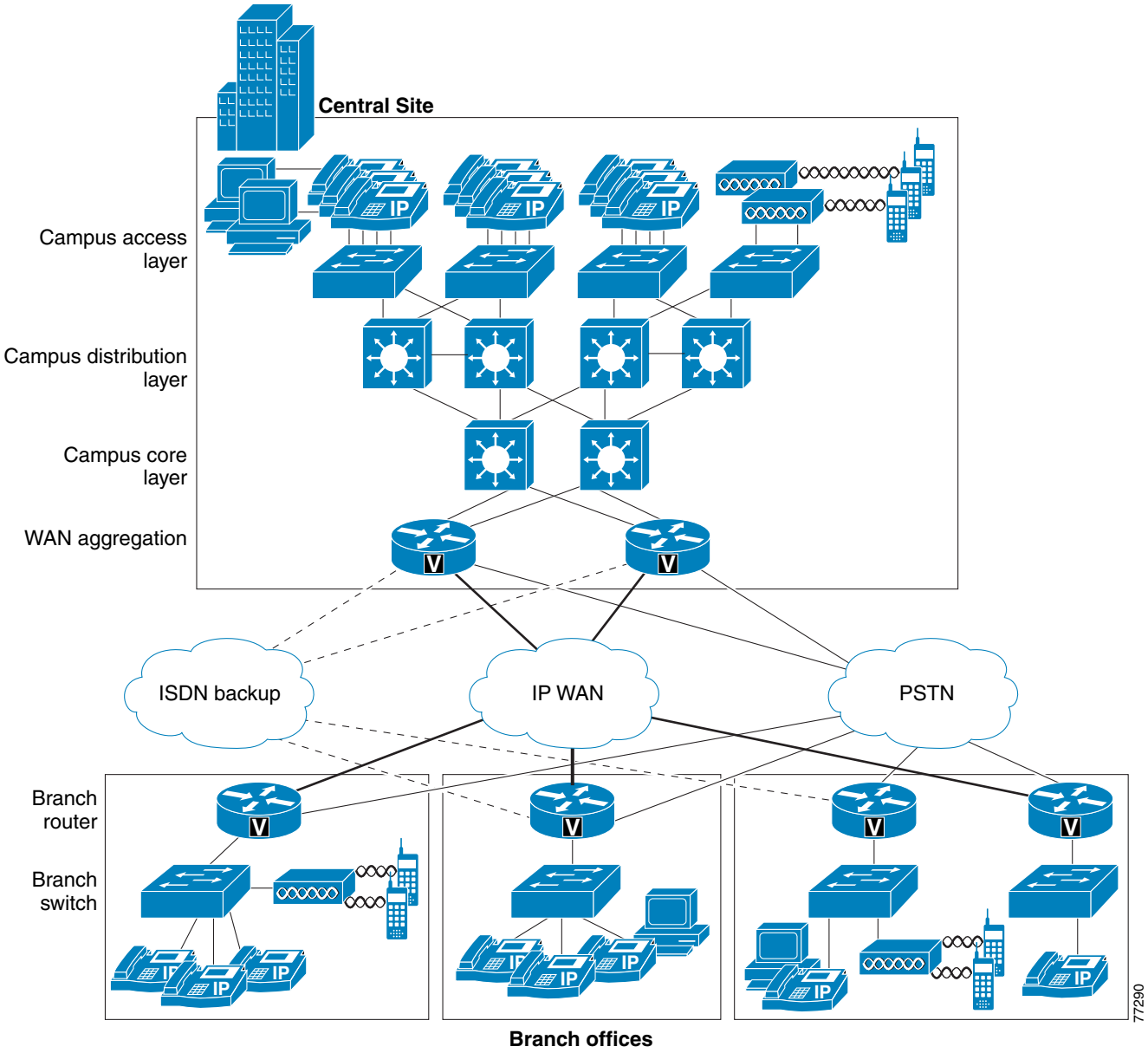


Table 3-1 Required Features for Each Role in the Network Infrastructure

Infrastructure Role	Required Features
Campus Access Switch	• In-Line Power[1] • Multiple Queue Support • 802.1p and 802.1Q • Fast Link Convergence
Campus Distribution or Core Switch	• Multiple Queue Support • 802.1p and 802.1Q • Traffic Classification • Traffic Reclassification
WAN Aggregation Router (Site that is at the hub of the network)	• Multiple Queue Support • Traffic Shaping • Link Fragmentation and Interleaving (LFI)[2] • Link Efficiency • Traffic Classification • Traffic Reclassification • 802.1p and 802.1Q
Branch Router (Spoke site)	• Multiple Queue Support • LFI[2] • Link Efficiency • Traffic Classification • Traffic Reclassification • 802.1p and 802.1Q
Branch or Smaller Site Switch	• In-Line Power[1] • Multiple Queue Support • 802.1p and 802.1Q

1. Recommended.
2. For link speeds less than 786 kbps.

What's New in This Chapter

Table 3-2 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 3-2 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Bandwidth provisioning for call control traffic	Provisioning for Call Control Traffic with Centralized Call Processing, page 3-57	March 1, 2018
Cisco Nexus 1000V Switch has been removed from this chapter	No longer in this document	March 1, 2018

LAN Infrastructure

Campus LAN infrastructure design is extremely important for proper Unified Communications operation on a converged network. Proper LAN infrastructure design requires following basic configuration and design best practices for deploying a highly available network. Further, proper LAN infrastructure design requires deploying end-to-end QoS on the network. The following sections discuss these requirements:

- LAN Design for High Availability, page 3-4
- LAN Quality of Service (QoS), page 3-14

LAN Design for High Availability

Properly designing a LAN requires building a robust and redundant network from the top down. By structuring the LAN as a layered model (see Figure 3-1) and developing the LAN infrastructure one step of the model at a time, you can build a highly available, fault tolerant, and redundant network. Once these layers have been designed correctly, you can add network services such as DHCP and TFTP to provide additional network functionality. The following sections examine the infrastructure layers and network services:

- Campus Access Layer, page 3-4
- Campus Distribution Layer, page 3-9
- Campus Core Layer, page 3-11
- Network Services, page 3-23

For more information on campus design, refer to the *Design Zone for Campus* at

https://www.cisco.com/go/designzone

Campus Access Layer

The access layer of the Campus LAN includes the portion of the network from the desktop port(s) to the wiring closet switch. Access layer switches have traditionally been configured as Layer 2 devices with Layer 2 uplinks to the distribution layer. The Layer 2 and spanning tree recommendations for Layer 2 access designs are well documented and are discussed briefly below. For newer Cisco Catalyst switches

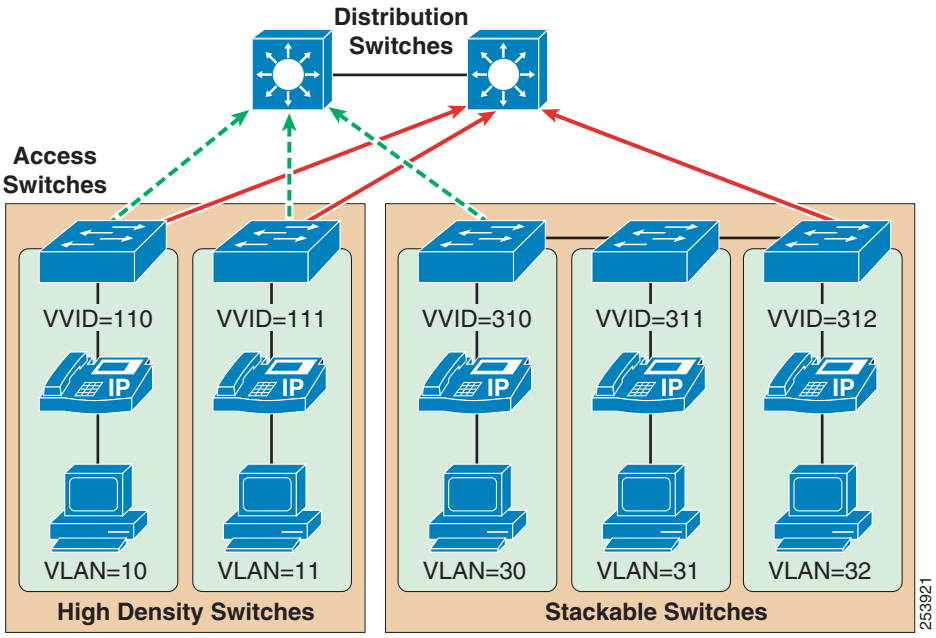
supporting Layer 3 protocols, new routed access designs are possible and offer improvements in convergence times and design simplicity. Routed access designs are discussed in the section on Routed Access Layer Designs, page 3-7.

Layer 2 Access Design Recommendations

Proper access layer design starts with assigning a single IP subnet per virtual LAN (VLAN). Typically, a VLAN should not span multiple wiring closet switches; that is, a VLAN should have presence in one and only one access layer switch (see Figure 3-2). This practice eliminates topological loops at Layer 2, thus avoiding temporary flow interruptions due to Spanning Tree convergence. However, with the introduction of standards-based IEEE 802.1w Rapid Spanning Tree Protocol (RSTP) and 802.1s Multiple Instance Spanning Tree Protocol (MISTP), Spanning Tree can converge at much higher rates. More importantly, confining a VLAN to a single access layer switch also serves to limit the size of the broadcast domain. There is the potential for large numbers of devices within a single VLAN or broadcast domain to generate large amounts of broadcast traffic periodically, which can be problematic. A good rule of thumb is to limit the number of devices per VLAN to about 512, which is equivalent to two Class C subnets (that is, a 23-bit subnet masked Class C address). For more information on the campus access layer, refer to the documentation on available at https://www.cisco.com/en/US/products/hw/switches/index.html.

Note The recommendation to limit the number of devices in a single Unified Communications VLAN to approximately 512 is not solely due to the need to control the amount of VLAN broadcast traffic. Installing Unified CM in a VLAN with an IP subnet containing more than 1024 devices can cause the Unified CM server ARP cache to fill up quickly, which can seriously affect communications between the Unified CM server and other Unified Communications endpoints.

Figure 3-2 Access Layer Switches and VLANs for Voice and Data



When you deploy voice, Cisco recommends that you enable two VLANs at the access layer: a native VLAN for data traffic (VLANs 10, 11, 30, 31, and 32 in Figure 3-2) and a voice VLAN under Cisco IOS or Auxiliary VLAN under CatOS for voice traffic (represented by VVIDs 110, 111, 310, 311, and 312 in Figure 3-2).

Separate voice and data VLANs are recommended for the following reasons:

- Address space conservation and voice device protection from external networks

  Private addressing of phones on the voice or auxiliary VLAN ensures address conservation and ensures that phones are not accessible directly through public networks. PCs and servers are typically addressed with publicly routed subnet addresses; however, voice endpoints may be addressed using RFC 1918 private subnet addresses.

- QoS trust boundary extension to voice and video devices

  QoS trust boundaries can be extended to voice and video devices without extending these trust boundaries and, in turn, QoS features to PCs and other data devices. For more information on trusted and untrusted devices, see the chapter on Bandwidth Management, page 13-1.

- Protection from malicious network attacks

  VLAN access control, 802.1Q, and 802.1p tagging can provide protection for voice devices from malicious internal and external network attacks such as worms, denial of service (DoS) attacks, and attempts by data devices to gain access to priority queues through packet tagging.

- Ease of management and configuration

  Separate VLANs for voice and data devices at the access layer provide ease of management and simplified QoS configuration.

To provide high-quality voice and to take advantage of the full voice feature set, access layer switches should provide support for:

- 802.1Q trunking and 802.1p for proper treatment of Layer 2 CoS packet marking on ports with phones connected
- Multiple egress queues to provide priority queuing of RTP voice packet streams
- The ability to classify or reclassify traffic and establish a network trust boundary
- Inline power capability (Although inline power capability is not mandatory, it is highly recommended for the access layer switches.)
- Layer 3 awareness and the ability to implement QoS access control lists (These features are recommended if you are using certain Unified Communications endpoints such as a PC running a softphone application like Jabber that cannot benefit from an extended trust boundary.)

Spanning Tree Protocol (STP)

To minimize convergence times and maximize fault tolerance at Layer 2, enable the following STP features:

- PortFast

  Enable PortFast on all access ports. The phones, PCs, or servers connected to these ports do not forward bridge protocol data units (BPDUs) that could affect STP operation. PortFast ensures that the phone or PC, when connected to the port, is able to begin receiving and transmitting traffic immediately without having to wait for STP to converge.

- Root guard or BPDU guard

  Enable root guard or BPDU guard on all access ports to prevent the introduction of a rogue switch that might attempt to become the Spanning Tree root, thereby causing STP re-convergence events and potentially interrupting network traffic flows. Ports that are set to **errdisable** state by BPDU guard must either be re-enabled manually or the switch must be configured to re-enable ports automatically from the errdisable state after a configured period of time.

- UplinkFast and BackboneFast

  Enable these features where appropriate to ensure that, when changes occur on the Layer 2 network, STP converges as rapidly as possible to provide high availability. When using Cisco stackable switches, enable Cross-Stack UplinkFast (CSUF) to provide fast failover and convergence if a switch in the stack fails.

- UniDirectional Link Detection (UDLD)

  Enable this feature to reduce convergence and downtime on the network when link failures or misbehaviors occur, thus ensuring minimal interruption of network service. UDLD detects, and takes out of service, links where traffic is flowing in only one direction. This feature prevents defective links from being mistakenly considered as part of the network topology by the Spanning Tree and routing protocols.

Note With the introduction of RSTP 802.1w, features such as PortFast and UplinkFast are not required because these mechanisms are built in to this standard. If RSTP has been enabled on the Catalyst switch, these commands are not necessary.

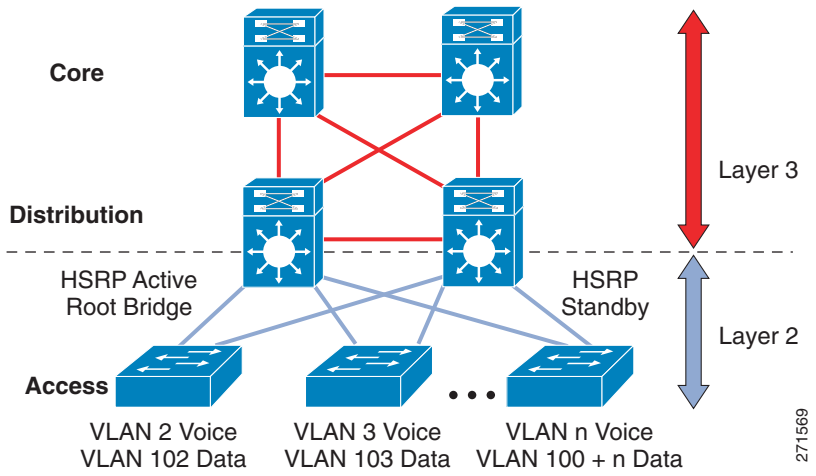
Routed Access Layer Designs

For campus designs requiring simplified configuration, common end-to-end troubleshooting tools, and the fastest convergence, a hierarchical design using Layer 3 switching in the access layer (routed access) in combination with Layer 3 switching at the distribution layer provides the fastest restoration of voice and data traffic flows.

Migrating the L2/L3 Boundary to the Access Layer

In the typical hierarchical campus design, the distribution layer uses a combination of Layer 2, Layer 3, and Layer 4 protocols and services to provide for optimal convergence, scalability, security, and manageability. In the most common distribution layer configurations, the access switch is configured as a Layer 2 switch that forwards traffic on high-speed trunk ports to the distribution switches. The distribution switches are configured to support both Layer 2 switching on their downstream access switch trunks and Layer 3 switching on their upstream ports toward the core of the network, as shown in Figure 3-3.

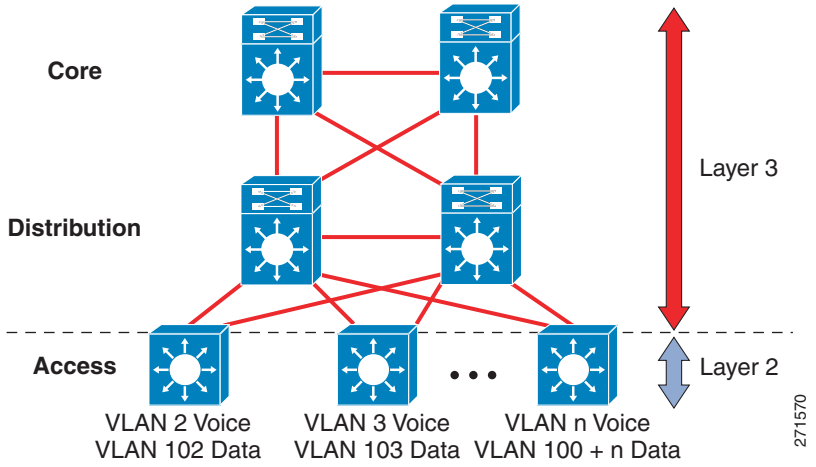
Figure 3-3 Traditional Campus Design — Layer 2 Access with Layer 3 Distribution



The purpose of the distribution switch in this design is to provide boundary functions between the bridged Layer 2 portion of the campus and the routed Layer 3 portion, including support for the default gateway, Layer 3 policy control, and all the multicast services required.

An alternative configuration to the traditional distribution layer model illustrated in Figure 3-3 is one in which the access switch acts as a full Layer 3 routing node (providing both Layer 2 and Layer 3 switching) and the access-to-distribution Layer 2 uplink trunks are replaced with Layer 3 point-to-point routed links. This alternative configuration, in which the Layer 2/3 demarcation is moved from the distribution switch to the access switch (as shown in Figure 3-4), appears to be a major change to the design but is actually just an extension of the current best-practice design.

Figure 3-4 Routed Access Campus Design — Layer 3 Access with Layer 3 Distribution



In both the traditional Layer 2 and the Layer 3 routed access designs, each access switch is configured with unique voice and data VLANs. In the Layer 3 design, the default gateway and root bridge for these VLANs is simply moved from the distribution switch to the access switch. Addressing for all end stations and for the default gateway remains the same. VLAN and specific port configurations remain

unchanged on the access switch. Router interface configuration, access lists, "ip helper," and any other configuration for each VLAN remain identical but are configured on the VLAN Switched Virtual Interface (SVI) defined on the access switch instead of on the distribution switches.

There are several notable configuration changes associated with the move of the Layer 3 interface down to the access switch. It is no longer necessary to configure a Hot Standby Router Protocol (HSRP) or Gateway Load Balancing Protocol (GLBP) virtual gateway address as the "router" interfaces because all the VLANs are now local. Similarly, with a single multicast router, for each VLAN it is not necessary to perform any of the traditional multicast tuning such as tuning PIM query intervals or ensuring that the designated router is synchronized with the active HSRP gateway.

Routed Access Convergence

The many potential advantages of using a Layer 3 access design include the following:

- Improved convergence
- Simplified multicast configuration
- Dynamic traffic load balancing
- Single control plane
- Single set of troubleshooting tools (for example, ping and traceroute)

Of these advantages, perhaps the most significant is the improvement in network convergence times possible when using a routed access design configured with Enhanced Interior Gateway Routing Protocol (EIGRP) or Open Shortest Path First (OSPF) as the routing protocol. Comparing the convergence times for an optimal Layer 2 access design (either with a spanning tree loop or without a loop) against that of the Layer 3 access design, you can obtain a four-fold improvement in convergence times, from 800 to 900 msec for the Layer 2 design to less than 200 msec for the Layer 3 access design.

For more information on routed access designs, refer to the document on *High Availability Campus Network Design – Routed Access Layer using EIGRP or OSPF*, available at

https://www.cisco.com/application/pdf/en/us/guest/netsol/ns432/c649/ccmigration_09186a0080811468.pdf

Campus Distribution Layer

The distribution layer of the Campus LAN includes the portion of the network from the wiring closet switches to the next-hop switch. For more information on the campus distribution layer switches, refer to the product documentation available at

https://www.cisco.com/en/US/products/hw/switches/index.html

At the distribution layer, it is important to provide redundancy to ensure high availability, including redundant links between the distribution layer switches (or routers) and the access layer switches. To avoid creating topological loops at Layer 2, use Layer 3 links for the connections between redundant Distribution switches when possible.

First-Hop Redundancy Protocols

In the campus hierarchical model, where the distribution switches are the L2/L3 boundary, they also act as the default gateway for the entire L2 domain that they support. Some form of redundancy is required because this environment can be large and a considerable outage could occur if the device acting as the default gateway fails.

Gateway Load Balancing Protocol (GLBP), Hot Standby Router Protocol (HSRP), and Virtual Router Redundancy Protocol (VRRP) are all first-hop redundancy protocols. Cisco initially developed HSRP to address the need for default gateway redundancy. The Internet Engineering Task Force (IETF) subsequently ratified Virtual Router Redundancy Protocol (VRRP) as the standards-based method of providing default gateway redundancy. More recently, Cisco developed GLBP to overcome some the limitations inherent in both HSRP and VRRP.

HSRP and VRRP with Cisco enhancements both provide a robust method of backing up the default gateway, and they can provide failover in less than one second to the redundant distribution switch when tuned properly.

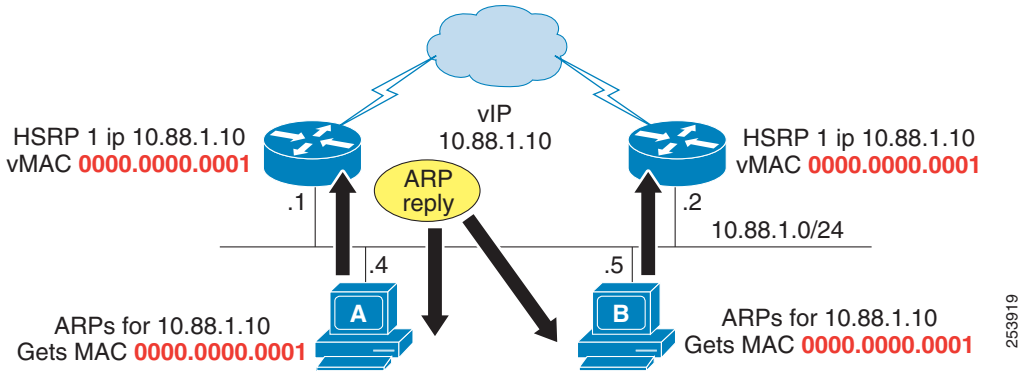
Gateway Load Balancing Protocol (GLBP)

Like HSRP and VRRP, Cisco's Gateway Load Balancing Protocol (GLBP) protects data traffic from a failed router or circuit, while also allowing packet load sharing between a group of redundant routers. When HSRP or VRRP are used to provide default gateway redundancy, the backup members of the peer relationship are idle, waiting for a failure event to occur for them to take over and actively forward traffic.

Before the development of GLBP, methods to utilize uplinks more efficiently were difficult to implement and manage. In one technique, the HSRP and STP/RSTP root alternated between distribution node peers, with the even VLANs homed on one peer and the odd VLANs homed on the alternate. Another technique used multiple HSRP groups on a single interface and used DHCP to alternate between the multiple default gateways. These techniques worked but were not optimal from a configuration, maintenance, or management perspective.

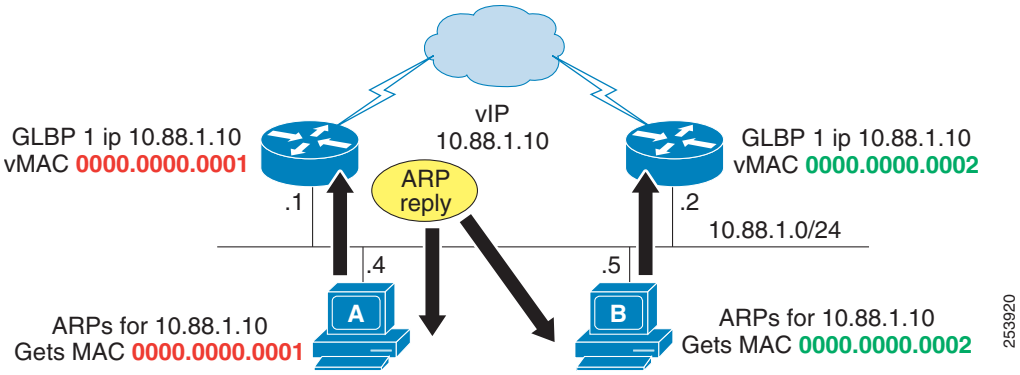
GLBP is configured and functions like HSRP. For HSRP, a single virtual MAC address is given to the endpoints when they use Address Resolution Protocol (ARP) to learn the physical MAC address of their default gateways (see Figure 3-5).

Figure 3-5 HSRP Uses One Virtual MAC Address



Two virtual MAC addresses exist with GLBP, one for each GLBP peer (see Figure 3-6). When an endpoint uses ARP to determine its default gateway, the virtual MAC addresses are checked in a round-robin basis. Failover and convergence work just like with HSRP. The backup peer assumes the virtual MAC address of the device that has failed, and begins forwarding traffic for its failed peer.

Figure 3-6 GLBP Uses Two Virtual MAC Addresses, One for Each GLBP Peer



The end result is that a more equal utilization of the uplinks is achieved with minimal configuration. As a side effect, a convergence event on the uplink or on the primary distribution node affects only half as many hosts, giving a convergence event an average of 50 percent less impact.

For more information on HSRP, VRRP, and GLBP, refer to the *Campus Network for High Availability Design Guide*, available at

https://www.cisco.com/application/pdf/en/us/guest/netsol/ns431/c649/ccmigration_09186a008093b876.pdf

Routing Protocols

Configure Layer 3 routing protocols such as OSPF and EIGRP at the distribution layer to ensure fast convergence, load balancing, and fault tolerance. Use parameters such as routing protocol timers, path or link costs, and address summaries to optimize and control convergence times as well as to distribute traffic across multiple paths and devices. Cisco also recommends using the **passive-interface** command to prevent routing neighbor adjacencies via the access layer. These adjacencies are typically unnecessary, and they create extra CPU overhead and increased memory utilization because the routing protocol keeps track of them. By using the **passive-interface** command on all interfaces facing the access layer, you prevent routing updates from being sent out on these interfaces and, therefore, neighbor adjacencies are not formed.

Campus Core Layer

The core layer of the Campus LAN includes the portion of the network from the distribution routers or Layer 3 switches to one or more high-end core Layer 3 switches or routers. Layer 3-capable Catalyst switches at the core layer can provide connectivity between numerous campus distribution layers. For more details on the campus core layer switches, refer to the documentation on available at https://www.cisco.com/en/US/products/hw/switches/index.html.

At the core layer, it is again very important to provide the following types of redundancy to ensure high availability:

- Redundant link or cable paths

  Redundancy here ensures that traffic can be rerouted around downed or malfunctioning links.

- Redundant devices

  Redundancy here ensures that, in the event of a device failure, another device in the network can continue performing tasks that the failed device was doing.

- Redundant device sub-systems

  This type of redundancy ensures that multiple power supplies and modules are available within a device so that the device can continue to function in the event that one of these components fails.

The Cisco Catalyst switches with Virtual Switching System (VSS) is a method to ensure redundancy in all of these areas by pooling together two Catalyst supervisor engines to act as one. For more information regarding VSS, refer to the product documentation available at

https://www.cisco.com/en/US/products/ps9336/index.html

Routing protocols at the core layer should again be configured and optimized for path redundancy and fast convergence. There should be no STP in the core because network connectivity should be routed at Layer 3. Finally, each link between the core and distribution devices should belong to its own VLAN or subnet and be configured using a 30-bit subnet mask.

Data Center and Server Farm

Typically, Cisco Unified Communications Manager (Unified CM) cluster servers, including media resource servers, reside in a firewall-secured data center or server farm environment. In addition, centralized gateways and centralized hardware media resources such as conference bridges, DSP or transcoder farms, and media termination points may be located in the data center or server farm. The placement of firewalls in relation to Cisco Unified Communications Manager (Unified CM) cluster servers and media resources can affect how you design and implement security in your network. For design guidance on firewall placement in relation to Unified Communications systems and media resources, see Firewalls, page 4-33.

Because these servers and resources are critical to voice networks, Cisco recommends distributing all Unified CM cluster servers, centralized voice gateways, and centralized hardware resources between multiple physical switches and, if possible, multiple physical locations within the campus. This distribution of resources ensures that, given a hardware failure (such as a switch or switch line card failure), at least some servers in the cluster will still be available to provide telephony services. In addition, some gateways and hardware resources will still be available to provide access to the PSTN and to provide auxiliary services. Besides being physically distributed, these servers, gateways, and hardware resources should be distributed among separate VLANs or subnets so that, if a broadcast storm or denial of service attack occurs on a particular VLAN, not all voice connectivity and services will be disrupted.

Power over Ethernet (PoE)

PoE (or inline power) is 48 Volt DC power provided over standard Ethernet unshielded twisted-pair (UTP) cable. Instead of using wall power, IP phones and other inline powered devices (PDs) such as the Aironet Wireless Access Points can receive power provided by inline power-capable Catalyst Ethernet switches or other inline power source equipment (PSE). Inline power is enabled by default on all inline power-capable Catalyst switches.

Deploying inline power-capable switches with uninterruptible power supplies (UPS) ensures that IP phones continue to receive power during power failure situations. Provided the rest of the telephony network is available during these periods of power failure, then IP phones should be able to continue making and receiving calls. You should deploy inline power-capable switches at the campus access layer within wiring closets to provide inline-powered Ethernet ports for IP phones, thus eliminating the need for wall power.

Caution The use of power injectors or power patch panels to deliver PoE can damage some devices because power is always applied to the Ethernet pairs. PoE switch ports automatically detect the presence of a device that requires PoE before enabling it on a port-by-port basis.

In addition to Cisco PoE inline power, Cisco now supports the IEEE 802.3af PoE and the IEEE 802.3at Enhanced PoE standards. For information on which Cisco Unified IP Phones support the 802.3af and 802.3at standards, refer to the product documentation for your particular phone models.

Energy Conservation for IP Phones

Cisco EnergyWise Technology provides intelligent management of energy usage for devices on the IP network, including Unified Communications endpoints that use Power over Ethernet (PoE). Cisco EnergyWise architecture can turn power on and off to devices connected with PoE on EnergyWise enabled switches, based on a configurable schedule. For more information on EnergyWise, refer to the documentation at

https://www.cisco.com/en/US/products/ps10195/index.html

When the PoE switch powers off IP phones for EnergyWise conservation, the phones are completely powered down. EnergyWise shuts down inline power on the ports that connect to IP phones and does so by a schedule or by commands from network management tools. When power is disabled, no verification occurs to determine whether a phone has an active call. The power is turned off and any active call is torn down. The IP phone loses registration from Cisco Unified Communications Manager and no calls can be made to or from the phone. There is no mechanism on the phone to power it on, therefore emergency calling will not be available on that phone.

The IP phone can be restarted only when the switch powers it on again. After power is restored, the IP phones will reboot and undergo a recovery process that includes requesting a new IP address, downloading a configuration file, applying any new configuration parameters, downloading new firmware or locales, and registering with Cisco Unified CM.

The EnergyWise schedule is configured and managed on the Cisco Network Infrastructure. It does not require any configuration on the IP phone or on Cisco Unified CM. However, power consumption on the phone can also be managed by a device profile configured on Unified CM. The energy saving options provided by Unified CM include the following:

- Power Save Plus Mode, page 3-13
- Power Save Mode, page 3-14

Power Save Plus Mode

In Power Save Plus mode, the phone on and off times and the idle timeout periods can be configured on the IP phones. The Cisco IP Phones' EnergyWise Power Save Plus configuration options specify the schedule for the IP phones to sleep (power down) and wake (power up). This mode requires an EnergyWise enabled network. If EnergyWise is enabled, then the sleep and wake times, as well as other parameters, can be used to control power to the phones. The Power Save Plus parameters are configured in the product-specific device profile in Cisco Unified CM Administration and sent to the IP phones as part of the phone configuration XML file.

During the configured power off period in this power saving mode, the IP phone sends a request to the switch asking for a wake-up at a specified time. If the switch is EnergyWise enabled, it accepts the request and reduces the power to the phone port, putting the phone to sleep. The sleep mode reduces the power consumption of the phone to 1 watt or less. The phone is not completely powered off in this case. When the phone is sleeping, the PoE switch provides minimal power that illuminates the Select key on

the phone. A user can wake up the IP phone by using the Select button. The IP phone does not go into sleep mode if a call is active on the phone. Audio and visual alerts can optionally be configured to warn users before a phone enters the Power Save Plus mode. While the phone is in sleep mode, it is not registered to Cisco Unified CM and cannot receive any inbound calls. Use the Forward Unregistered setting in the phone's device configuration profile to specify how to treat any inbound calls to the phone's number.

Note The Cisco EnergyWise Power Save Plus mode is supported on most Cisco IP Phones and Collaboration Desk Endpoints. To learn which endpoints support EnergyWise Power Save Plus, refer to the data sheets for your endpoint models:
https://www.cisco.com/c/en/us/products/collaboration-endpoints/product-listing.html

Power Save Mode

In Power Save mode, the backlight on the screen is not lit when the phone is not in use. The phone stays registered to Cisco Unified CM in this mode and can receive inbound calls and make outbound calls. Cisco Unified CM Administration has product-specific configuration options to turn off the display at a designated time on some days and all day on other days. The phone remains in Power Save mode for the scheduled duration or until the user lifts the handset or presses any button. An EnergyWise enabled network is not required for the Power Save mode. Idle times can be scheduled so that the display remains on until the timeout and then turns off automatically. The phone is still powered on in this mode and can receive inbound calls.

The Power Save mode can be used together with the Power Save Plus mode. Using both significantly reduces the total power consumption by Cisco Unified IP Phones.

For information on configuring these modes, refer to the administration guides for the Cisco IP Phones and Collaboration Desk Endpoints:

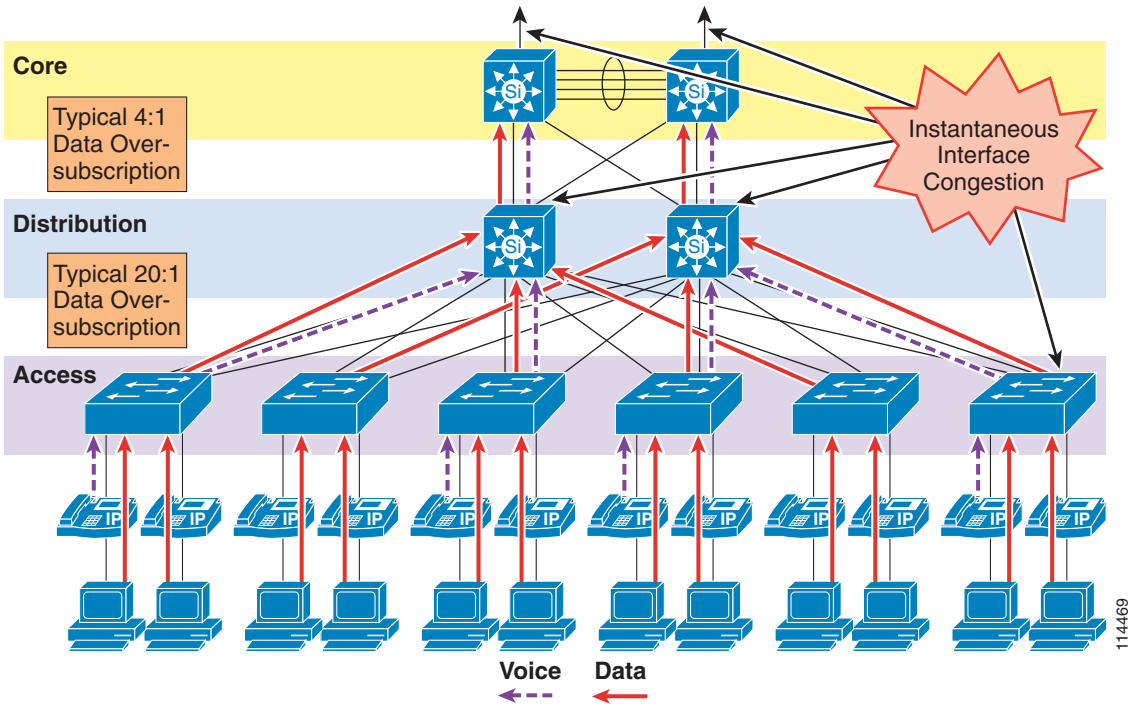
https://www.cisco.com/c/en/us/products/collaboration-endpoints/product-listing.html

LAN Quality of Service (QoS)

Until recently, quality of service was not an issue in the enterprise campus due to the asynchronous nature of data traffic and the ability of network devices to tolerate buffer overflow and packet loss. However, with new applications such as voice and video, which are sensitive to packet loss and delay, buffers and not bandwidth are the key QoS issue in the enterprise campus.

Figure 3-7 illustrates the typical oversubscription that occurs in LAN infrastructures.

Figure 3-7 Data Traffic Oversubscription in the LAN



This oversubscription, coupled with individual traffic volumes and the cumulative effects of multiple independent traffic sources, can result in the egress interface buffers becoming full instantaneously, thus causing additional packets to drop when they attempt to enter the egress buffer. The fact that campus switches use hardware-based buffers, which compared to the interface speed are much smaller than those found on WAN interfaces in routers, merely increases the potential for even short-lived traffic bursts to cause buffer overflow and dropped packets.

Applications such as file sharing (both peer-to-peer and server-based), remote networked storage, network-based backup software, and emails with large attachments, can create conditions where network congestion occurs more frequently and/or for longer durations. Some of the negative effects of recent worm attacks have been an overwhelming volume of network traffic (both unicast and broadcast-storm based), increasing network congestion. If no buffer management policy is in place, loss, delay, and jitter performance of the LAN may be affected for all traffic.

Another situation to consider is the effect of failures of redundant network elements, which cause topology changes. For example, if a distribution switch fails, all traffic flows will be reestablished through the remaining distribution switch. Prior to the failure, the load balancing design shared the load between two switches, but after the failure all flows are concentrated in a single switch, potentially causing egress buffer conditions that normally would not be present.

For applications such as voice, this packet loss and delay results in severe voice quality degradation. Therefore, QoS tools are required to manage these buffers and to minimize packet loss, delay, and delay variation (jitter).

The following types of QoS tools are needed end-to-end on the network to manage traffic and ensure voice and video quality:

- Traffic classification

  Classification involves the marking of packets with a specific priority denoting a requirement for class of service (CoS) from the network. The point at which these packet markings are trusted or not trusted is considered the trust boundary. Trust is typically extended to voice devices (phones) and not to data devices (PCs).

- Queuing or scheduling

  Interface queuing or scheduling involves assigning packets to one of several queues based on classification for expedited treatment throughout the network.

- Bandwidth provisioning

  Provisioning involves accurately calculating the required bandwidth for all applications plus element overhead.

The following sections discuss the use of these QoS mechanisms in a campus environment:

- Traffic Classification, page 3-16
- Interface Queuing, page 3-18
- Bandwidth Provisioning, page 3-19
- Impairments to IP Communications if QoS is Not Employed, page 3-19

Traffic Classification

It has always been an integral part of the Cisco network design architecture to classify or mark traffic as close to the edge of the network as possible. Traffic classification is an entrance criterion for access into the various queuing schemes used within the campus switches and WAN interfaces. Cisco IP Phones mark voice control signaling and voice RTP streams at the source, and they adhere to the values presented in Table 3-3. As such, the IP phone can and should classify traffic flows.

Table 3-3 lists the traffic classification requirements for the LAN infrastructure.

Table 3-3 Traffic Classification Guidelines for Various Types of Network Traffic

Application	Layer-3 Classification			Layer-2 Classification
	Type of Service (ToS) IP Precedence (IPP)	Per-Hop Behavior (PHB)	Differentiated Services Code Point (DSCP)	Class of Service (CoS)
Routing	6	CS6	48	6
Voice Real-Time Transport Protocol (RTP)	5	EF	46	5
Videoconferencing	4	AF41	34	4
IP video	4	AF41	34	4
Immersive video Real-Time Interactive	4	CS4	32	4
Streaming video	3	AF31	26	3
Call signaling	3	CS3	24	3
Transactional data	2	AF21	18	2
Network management	2	CS2	16	2
Scavenger	1	CS1	8	1
Best effort	0	0	0	0

For more information about traffic classification, refer to the QoS design guides available at

https://www.cisco.com/c/en/us/solutions/enterprise/design-zone-ipv6/design-guide-listing.html

Traffic Classification for Video Telephony

The main classes of interest for IP Video Telephony are:

- Voice

  Voice is classified as CoS 5 (IP Precedence 5, PHB EF, or DSCP 46).

- Videoconferencing

  Videoconferencing is classified as CoS 4 (IP Precedence 4, PHB AF41, or DSCP 34).

- Call signaling

  Call signaling for voice and videoconferencing is classified as CoS 3 (IP Precedence 3, PHB CS3, or DSCP 24).

Cisco highly recommends these classifications as *best practices* in a Cisco Unified Communications network.

QoS Marking Differences Between Video Calls and Voice-Only Calls

The voice component of a call can be classified in one of two ways, depending on the type of call in progress. A voice-only telephone call would have its media classified as CoS 5 (IP Precedence 5 or PHB EF), while the voice channel of a video conference would have its media classified as CoS 4 (IP Precedence 4 or PHB AF41). All the Cisco IP Video Telephony products adhere to the Cisco

Corporate QoS Baseline standard, which requires that the audio and video channels of a video call both be marked as CoS 4 (IP Precedence 4 or PHB AF41). The reasons for this recommendation include, but are not limited to, the following:

- To preserve lip-sync between the audio and video channels
- To provide separate classes for audio-only calls and video calls

Cisco is in the process of changing this requirement for endpoints to mark the audio and video channels of a video call separately, thus providing the flexibility to mark both the audio and video channels of a video call with the same DSCP value or different DSCP values, depending on the use cases. For more information on DSCP marking, see the chapter on Bandwidth Management, page 13-1.

The signaling class is applicable to all voice signaling protocols (such as SCCP, MGCP, and so on) as well as video signaling protocols (such as SCCP, H.225, RAS, CAST, and so on).

Given the recommended classes, the first step is to decide where the packets will be classified (that is, which device will be the first to mark the traffic with its QoS classification). There are essentially two places to mark or classify traffic:

- On the originating endpoint — the classification is then trusted by the upstream switches and routers
- On the switches and/or routers — because the endpoint is either not capable of classifying its own packets or is not trustworthy to classify them correctly

QoS Enforcement Using a Trusted Relay Point (TRP)

A Trusted Relay Point (TRP) can be used to enforce and/or re-mark the DSCP values of media flows from endpoints. This feature allows QoS to be enforced for media from endpoints such as softphones, where the media QoS values might have been modified locally.

A TRP is a media resource based upon the existing Cisco IOS media termination point (MTP) function.

Endpoints can be configured to "Use Trusted Relay Point," which will invoke a TRP for all calls.

For QoS enforcement, the TRP uses the configured QoS values for media in Unified CM's Service Parameters to re-mark and enforce the QoS values in media streams from the endpoint.

TRP functionality is supported by Cisco IOS MTPs and transcoding resources. (Use Unified CM to check "Enable TRP" on the MTP or transcoding resource to activate TRP functionality.)

Interface Queuing

After packets have been marked with the appropriate tag at Layer 2 (CoS) and Layer 3 (DSCP or PHB), it is important to configure the network to schedule or queue traffic based on this classification, so as to provide each class of traffic with the service it needs from the network. By enabling QoS on campus switches, you can configure all voice traffic to use separate queues, thus virtually eliminating the possibility of dropped voice packets when an interface buffer fills instantaneously.

Although network management tools may show that the campus network is not congested, QoS tools are still required to guarantee voice quality. Network management tools show only the average congestion over a sample time span. While useful, this average does not show the congestion peaks on a campus interface.

Transmit interface buffers within a campus tend to congest in small, finite intervals as a result of the bursty nature of network traffic. When this congestion occurs, any packets destined for that transmit interface are dropped. The only way to prevent dropped voice traffic is to configure multiple queues on campus switches. For this reason, Cisco recommends always using a switch that has at least two output queues on each port and the ability to send packets to these queues based on QoS Layer 2 and/or Layer 3

classification. The majority of Cisco Catalyst Switches support two or more output queues per port. For more information on Cisco Catalyst Switch interface queuing capabilities, refer to the documentation at https://www.cisco.com/en/US/products/hw/switches/index.html

Bandwidth Provisioning

In the campus LAN, bandwidth provisioning recommendations can be summarized by the motto, *Over provision and under subscribe*. This motto implies careful planning of the LAN infrastructure so that the available bandwidth is always considerably higher than the load and there is no steady-state congestion over the LAN links.

The addition of voice traffic onto a converged network does not represent a significant increase in overall network traffic load; the bandwidth provisioning is still driven by the demands of the data traffic requirements. The design goal is to avoid extensive data traffic congestion on any link that will be traversed by telephony signaling or media flows. Contrasting the bandwidth requirements of a single G.711 voice call (approximately 86 kbps) to the raw bandwidth of a FastEthernet link (100 Mbps) indicates that voice is not a source of traffic that causes network congestion in the LAN, but rather it is a traffic flow to be protected from LAN network congestion.

Impairments to IP Communications if QoS is Not Employed

If QoS is not deployed, packet drops and excessive delay and jitter can occur, leading to impairments of the telephony services. When media packets are subjected to drops, delay, and jitter, the user-perceivable effects include clicking sound, harsh-sounding voice, extended periods of silence, and echo.

When signaling packets are subjected to the same conditions, user-perceivable impairments include unresponsiveness to user input (such as delay to dial tone), continued ringing upon answer, and double dialing of digits due to the user's belief that the first attempt was not effective (thus requiring hang-up and redial). More extreme cases can include endpoint re-initialization, call termination, and the spurious activation of SRST functionality at branch offices (leading to interruption of gateway calls).

These effects apply to all deployment models. However, single-site (campus) deployments tend to be less likely to experience the conditions caused by sustained link interruptions because the larger quantity of bandwidth typically deployed in LAN environments (minimum links of 100 Mbps) allows for some residual bandwidth to be available for the IP Communications system.

In any WAN-based deployment model, traffic congestion is more likely to produce sustained and/or more frequent link interruptions because the available bandwidth is much less than in a LAN (typically less than 2 Mbps), so the link is more easily saturated. The effects of link interruptions can impact the user experience, whether or not the voice media traverses the packet network, because signaling traffic between endpoints and the Unified CM servers can also be delayed or dropped.

QoS Design Considerations for Virtual Unified Communications with Cisco UCS Servers

Unified Communications applications such as Cisco Unified Communications Manager (Unified CM) run as virtual machines on top of the VMware Hypervisor. These Unified Communications virtual machines are connected to a virtual software switch rather than a hardware-based Ethernet. The following types of virtual software switches are available:

- VMware vSphere Standard Switch

  Available with all VMware vSphere editions and independent of the type of VMware licensing scheme. The vSphere Standard Switch exists only on the host on which it is configured.

- VMware vSphere Distributed Switch

  Available only with the Enterprise Plus Edition of VMware vSphere. The vSphere Distributed Switch acts as a single switch across all associated hosts on a datacenter and helps simplify manageability of the software virtual switch.

From the point of view of virtual connectivity, each virtual machine can connect to any one of the above virtual switches residing on a blade server. When using Cisco UCS B-Series blade servers, the blade servers physically connect to the rest of the network through a Fabric Extender in the UCS chassis to a UCS Fabric Interconnect Switch (for example, Cisco UCS 6200 Series). The UCS Fabric Interconnect Switch is where the physical wiring connects to a customer's Ethernet LAN and FC SAN.

From the point of view of traffic flow, traffic from the virtual machines first goes to the software virtual switch (for example, vSphere Standard Switch or vSphere Distributed Switch). The virtual switch then sends the traffic to the physical UCS Fabric Interconnect Switch through its blade server's Network Adapter and Fabric Extender. The UCS Fabric Interconnect Switch carries both the IP and fibre channel SAN traffic via Fibre Channel over Ethernet (FCoE) on a single wire. The UCS Fabric Interconnect Switch sends IP traffic to an IP switch (for example, Cisco Catalyst or Nexus Series Switch), and it sends SAN traffic to a Fibre Channel SAN Switch (for example, Cisco MDS Series Switch).

Congestion Scenario

In a deployment with Cisco UCS B-Series blades servers and with Cisco Collaboration applications only, network congestion or an oversubscription scenario is unlikely because the UCS Fabric Interconnect Switch provides a high-capacity switching fabric, and the usable bandwidth per server blade far exceeds the maximum traffic requirements of a typical Collaboration application.

However, there might be scenarios where congestion could arise. For example, with a large number of B-Series blade servers and chassis, a large number of applications, and/or third-party applications requiring high network bandwidth, there is a potential for congestion on the different network elements of the UCS B-Series system (adapters, IO modules, Fabric Interconnects). In addition, FCoE traffic is sharing the same network elements as IP traffic, therefore applications performing a high amount of storage transfer would increase the utilization on the network elements and contribute to this potential congestion.

To address this potential congestion, QoS should be implemented.

QoS Implementation with Cisco UCS B-Series

Cisco UCS Fabric Interconnect Switches and adapters such as the Cisco VIC adapter perform QoS based on Layer 2 CoS values. Traffic types are classified by CoS value into QoS system classes that determine, for example, the minimum amount of bandwidth guaranteed and the packet drop policy to be used for

each class. However, Cisco Collaboration applications perform QoS marking at Layer 3 only, not at the Layer 2. Hence the need for mapping the L3 values used by the applications to the L2 CoS values used by the Cisco UCS elements.

The VMware vSphere Standard Switch, vSphere Distributed Switch, Cisco UCS Fabric Interconnect switches, and other UCS network elements do not have the ability to perform this mapping between L3 and L2 values.

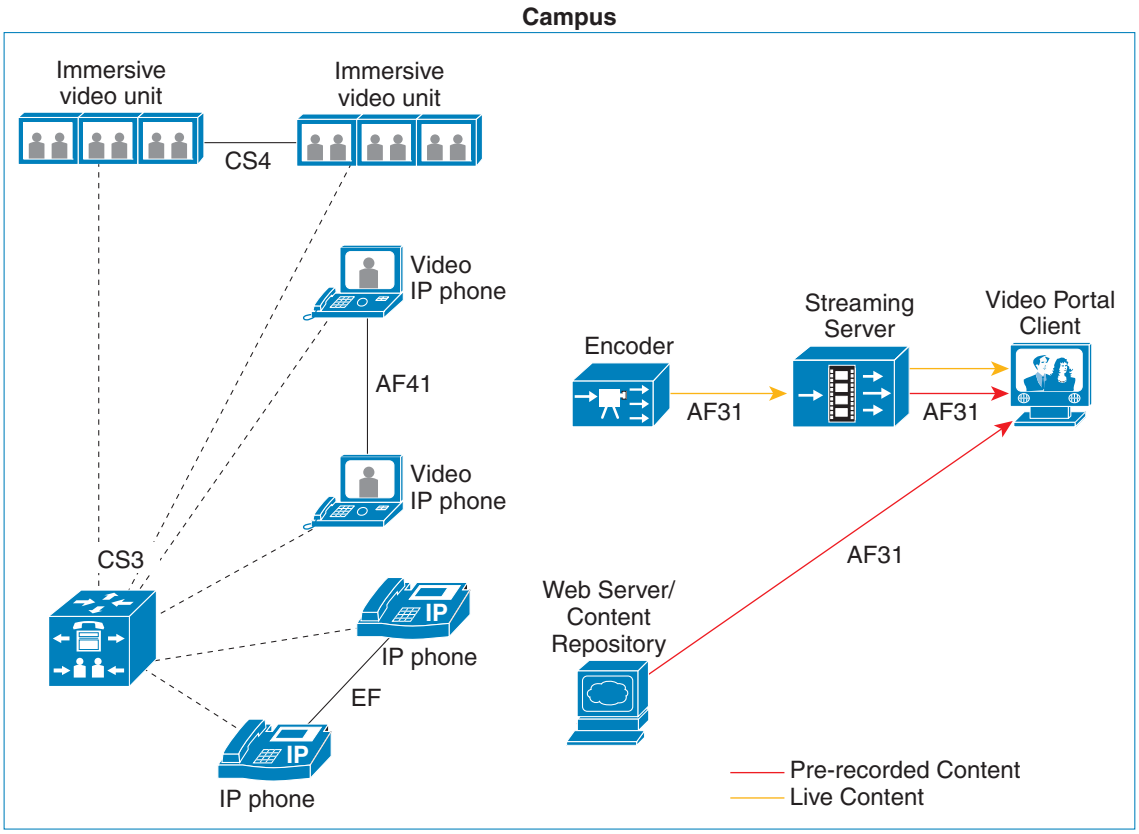
Note Fibre Channel over Ethernet (FCoE) traffic has a reserved QoS system class that should not be used by any other type of traffic. By default, this system class has a CoS value of 3, which is the same value assigned to the system class used by voice and video signaling traffic in the example above. To prevent voice and video signaling traffic from using the FCoE system class, assign a different CoS value to the FCoE system class (2 or 4, for instance).

To work around this issue, you could create multiple virtual switches and assign a different CoS value for the uplink ports of each of those switches. For example, virtual switch 1 would have uplink ports configured with a CoS value of 1, virtual switch 2 would have uplink ports configured with a CoS value of 2, and so forth. Then the application virtual machines would be assigned to a virtual switch, depending on the desired QoS system class. The downside to this approach is that all traffic types from a virtual machine will have the same CoS value. For example, with a Unified CM virtual machine, real-time media traffic such as MoH traffic, signaling traffic, and non-voice traffic (for example, backups, CDRs, logs, Web traffic, and so forth) would share the same CoS value.

QoS Design Considerations for Video

Cisco recommends using different DSCP markings for different video applications. Unified CM 9.*x* provides support for different DSCP markings for immersive video traffic and videoconferencing (IP video telephony) traffic. By default, Unified CM 9.*x* has preconfigured the recommended DSCP values for TelePresence (immersive video) calls at CS4 and video (IP video telephony) calls at AF41. Figure 3-8 depicts the different video applications in a converged environment using the recommended DSCP values.

Figure 3-8 Recommended QoS Traffic Markings in a Converged Network



Calculating Overhead for QoS

Unlike voice, real-time IP video traffic in general is a somewhat bursty, variable bit rate stream. Therefore video, unlike voice, does not have clear formulas for calculating network overhead because video packet sizes and rates vary proportionally to the degree of motion within the video image itself. From a network administrator's point of view, bandwidth is always provisioned at Layer 2, but the variability in the packet sizes and the variety of Layer 2 media that the packets may traverse from end-to-end make it difficult to calculate the real bandwidth that should be provisioned at Layer 2. However, the conservative rule that has been thoroughly tested and widely used is to over-provision video bandwidth by 20%. This accommodates the 10% burst and the network overhead from Layer 2 to Layer 4.

Network Services

The deployment of an IP Communications system requires the coordinated design of a well structured, highly available, and resilient network infrastructure as well as an integrated set of network services including Domain Name System (DNS), Dynamic Host Configuration Protocol (DHCP), Trivial File Transfer Protocol (TFTP), and Network Time Protocol (NTP).

Domain Name System (DNS)

DNS enables the mapping of host names and network services to IP addresses within a network or networks. DNS server(s) deployed within a network provide a database that maps network services to hostnames and, in turn, hostnames to IP addresses. Devices on the network can query the DNS server and receive IP addresses for other devices in the network, thereby facilitating communication between network devices.

A complete collaboration solution relies on DNS in order to function correctly for a number of services and thus requires a highly available DNS structure in place. For basic IP telephony deployments where reliance on DNS is not desired, Unified CM can be configured to support and ensure communication between Unified CM(s), gateways, and endpoint devices using IP addresses rather than hostnames.

Deploying Unified CM without DNS

For basic IP telephony deployments where DNS is not desired, Cisco recommends that you configure Unified CM(s), gateways, and endpoint devices to use IP addresses rather than hostnames. This should be done during installation of the Unified CM cluster. During installation of the publisher and subscriber nodes, Cisco recommends that you do not select the option to enable DNS. After the initial installation of the publisher node in a Unified CM cluster, the publisher will be referenced in the server table by the hostname you provided for the system. Before installation and configuration of any subsequent subscriber nodes or the definition of any endpoints, you should change this server entry to the IP address of the publisher node rather than the hostname. Each subscriber node added to the cluster should be defined in this same server table by IP address and not by hostname. Each subscriber node should be added to this server table one device at a time, and there should be no definitions for non-existent subscriber nodes at any time other than for the new subscriber node being installed.

Deploying Unified CM with DNS

You should always deploy DNS servers in a geographically redundant fashion so that a single DNS server failure will not prevent network communications between IP telephony devices. By providing DNS server redundancy in the event of a single DNS server failure, you ensure that devices relying on DNS to communicate on the network can still receive hostname-to-IP-address mappings from a backup or secondary DNS server.

Unified CM can use DNS to:

- Provide simplified system management
- Resolve fully qualified domain names to IP addresses for trunk destinations
- Resolve fully qualified domain names to IP addresses for SIP route patterns based on domain name
- Resolve service (SRV) records to host names and then to IP addresses for SIP trunk destinations
- Provide certificate-based security

Collaboration clients use DNS for:

- Single Sign-On (SSO)
- Jabber deployments requiring user registration auto-discovery
- Certificate-based security for secure signaling and media

When DNS is used, Cisco recommends defining each Unified CM cluster as a member of a valid sub-domain within the larger organizational DNS domain, defining the DNS domain on each Cisco Unified CM server, and defining the primary and secondary DNS server addresses on each Unified CM server.

Table 3-4 shows an example of how DNS server could use A records (Hostname-to-IP-address resolution), Cname records (aliases), and SRV records (service records for redundancy, load balancing, and service discovery) in a Unified CM environment.

Table 3-4 Example Use of DNS with Unified CM

Host Name	Type	TTL	Data
CUCM-Admin.cluster1.cisco.com	Host (A)	12 Hours	182.10.10.1
CUCM1.cluster1.cisco.com	Host (A)	Default	182.10.10.1
CUCM2.cluster1.cisco.com	Host (A)	Default	182.10.10.2
CUCM3.cluster1.cisco.com	Host (A)	Default	182.10.10.3
CUCM4.cluster1.cisco.com	Host (A)	Default	182.10.10.4
TFTP-server1.cluster1.cisco.com	Host (A)	12 Hours	182.10.10.11
TFTP-server2.cluster1.cisco.com	Host (A)	12 Hours	182.10.10.12
CUP1.cluster1.cisco.com	Host (A)	Default	182.10.10.15
CUP2.cluster1.cisco.com	Host (A)	Default	182.10.10.16
www.CUCM-Admin.cisco.com	Alias (CNAME)	Default	CUCM-Admin.cluster1.cisco.com
_sip._tcp.cluster1.cisco.com.	Service (SRV)	Default	CUCM1.cluster1.cisco.com
_sip._tcp.cluster1.cisco.com.	Service (SRV)	Default	CUCM2.cluster1.cisco.com
_sip._tcp.cluster1.cisco.com.	Service (SRV)	Default	CUCM3.cluster1.cisco.com
_sip._tcp.cluster1.cisco.com.	Service (SRV)	Default	CUCM4.cluster1.cisco.com

For Jabber clients, refer to the *Cisco Jabber DNS Configuration Guide*, available at

https://www.cisco.com/web/products/voice/jabber.html

Dynamic Host Configuration Protocol (DHCP)

DHCP is used by hosts on the network to obtain initial configuration information, including IP address, subnet mask, default gateway, and TFTP server address. DHCP eases the administrative burden of manually configuring each host with an IP address and other configuration information. DHCP also provides automatic reconfiguration of network configuration when devices are moved between subnets. The configuration information is provided by a DHCP server located in the network, which responds to DHCP requests from DHCP-capable clients.

You should configure IP Communications endpoints to use DHCP to simplify deployment of these devices. Any RFC 2131 compliant DHCP server can be used to provide configuration information to IP Communications network devices. When deploying IP telephony devices in an existing data-only

network, all you have to do is add DHCP voice scopes to an existing DHCP server for these new voice devices. Because IP telephony devices are configured to use and rely on a DHCP server for IP configuration information, you must deploy DHCP servers in a redundant fashion. At least two DHCP servers should be deployed within the telephony network such that, if one of the servers fails, the other can continue to answer DHCP client requests. You should also ensure that DHCP server(s) are configured with enough IP subnet addresses to handle all DHCP-reliant clients within the network.

DHCP Option 150

IP telephony endpoints can be configured to rely on DHCP Option 150 to identify the source of telephony configuration information, available from a server running the Trivial File Transfer Protocol (TFTP).

In the simplest configuration, where a single TFTP server is offering service to all deployed endpoints, Option 150 is delivered as a single IP address pointing to the system's designated TFTP server. The DHCP scope can also deliver two IP addresses under Option 150, for deployments where there are two TFTP servers within the same cluster. The phone would use the second address if it fails to contact the primary TFTP server, thus providing redundancy. To achieve both redundancy and load sharing between the TFTP servers, you can configure Option 150 to provide the two TFTP server addresses in reverse order for half of the DHCP scopes.

Note If the primary TFTP server is available but is not able to grant the requested file to the phone (for example, because the requesting phone is not configured on that cluster), the phone will not attempt to contact the secondary TFTP server.

Cisco highly recommends using a direct IP address (that is, not relying on a DNS service) for Option 150 because doing so eliminates dependencies on DNS service availability during the phone boot-up and registration process.

Note Even though IP phones support a maximum of two TFTP servers under Option 150, you could configure a Unified CM cluster with more than two TFTP servers. For instance, if a Unified CM system is clustered over a WAN at three separate sites, three TFTP servers could be deployed (one at each site). Phones within each site could then be granted a DHCP scope containing that site's TFTP server within Option 150. This configuration would bring the TFTP service closer to the endpoints, thus reducing latency and ensuring failure isolation between the sites (one site's failure would not affect TFTP service at another site).

Phone DHCP Operation Following a Power Recycle

If a phone is powered down and comes back up while the DHCP server is still offline, it will attempt to use DHCP to obtain IP addressing information (as normal). In the absence of a response from a DHCP server, the phone will re-use the previously received DHCP information to register with Unified CM.

DHCP Lease Times

Configure DHCP lease times as appropriate for the network environment. Given a fairly static network in which PCs and telephony devices remain in the same place for long periods of time, Cisco recommends longer DHCP lease times (for example, one week). Shorter lease times require more frequent renewal of the DHCP configuration and increase the amount of DHCP traffic on the network. Conversely, networks that incorporate large numbers of mobile devices, such as laptops and wireless telephony devices, should be configured with shorter DHCP lease times (for example, one day) to

prevent depletion of DHCP-managed subnet addresses. Mobile devices typically use IP addresses for short increments of time and then might not request a DHCP renewal or new address for a long period of time. Longer lease times will tie up these IP addresses and prevent them from being reassigned even when they are no longer being used.

Cisco Unified IP Phones adhere to the conditions of the DHCP lease duration as specified in the DHCP server's scope configuration. Once half the lease time has expired since the last successful DHCP server acknowledgment, the IP phone will request a lease renewal. This DHCP client Request, once acknowledged by the DHCP server, will allow the IP phone to retain use of the IP scope (that is, the IP address, default gateway, subnet mask, DNS server (optional), and TFTP server (optional)) for another lease period. If the DHCP server becomes unavailable, an IP phone will not be able to renew its DHCP lease, and as soon as the lease expires, it will relinquish its IP configuration and will thus become unregistered from Unified CM until a DHCP server can grant it another valid scope.

In centralized call processing deployments, if a remote site is configured to use a centralized DHCP server (through the use of a DHCP relay agent such as the IP Helper Address in Cisco IOS) and if connectivity to the central site is severed, IP phones within the branch will not be able to renew their DHCP scope leases. In this situation, branch IP phones are at risk of seeing their DHCP lease expire, thus losing the use of their IP address, which would lead to service interruption. Given the fact that phones attempt to renew their leases at half the lease time, DHCP lease expiration can occur as soon as half the lease time since the DHCP server became unreachable. For example, if the lease time of a DHCP scope is set to 4 days and a WAN failure causes the DHCP server to be unavailable to the phones in a branch, those phones will be unable to renew their leases at half the lease time (in this case, 2 days). The IP phones could stop functioning as early as 2 days after the WAN failure, unless the WAN comes back up and the DHCP server is available before that time. If the WAN connectivity failure persists, all phones see their DHCP scope expire after a maximum of 4 days from the WAN failure.

This situation can be mitigated by one of the following methods:

- Set the DHCP scope lease to a long duration (for example, 8 days or more).

  This method would give the system administrator a minimum of half the lease time to remedy any DHCP reachability problem. Long lease durations also have the effect of reducing the frequency of network traffic associated with lease renewals.

- Configure co-located DHCP server functionality (for example, run a DHCP server function on the branch's Cisco IOS router).

  This approach is immune to WAN connectivity interruption. One effect of such an approach is to decentralize the management of IP addresses, requiring incremental configuration efforts in each branch. (See DHCP Network Deployments, page 3-26, for more information.)

Note The term *co-located* refers to two or more devices in the same physical location, with no WAN or MAN connection between them.

DHCP Network Deployments

There are two options for deploying DHCP functionality within an IP telephony network:

- Centralized DHCP Server

  Typically, for a single-site campus IP telephony deployment, the DHCP server should be installed at a central location within the campus. As mentioned previously, redundant DHCP servers should be deployed. If the IP telephony deployment also incorporates remote branch telephony sites, as in a centralized multisite Unified CM deployment, a centralized server can be used to provide DHCP service to devices in the remote sites. This type of deployment requires that you configure the **ip helper-address** on the branch router interface. Keep in mind that, if redundant DHCP servers are

deployed at the central site, both servers' IP addresses must be configured as **ip helper-address**. Also note that, if branch-side telephony devices rely on a centralized DHCP server and the WAN link between the two sites fails, devices at the branch site will be unable to send DHCP requests or receive DHCP responses.

Note By default, **service dhcp** is enabled on the Cisco IOS device and does not appear in the configuration. Do not disable this service on the branch router because doing so will disable the DHCP relay agent on the device, and the **ip helper-address** configuration command will not work.

- Centralized DHCP Server and Remote Site Cisco IOS DHCP Server

  When configuring DHCP for use in a centralized multisite Unified CM deployment, you can use a centralized DHCP server to provide DHCP service to centrally located devices. Remote devices could receive DHCP service from a locally installed server or from the Cisco IOS router at the remote site. This type of deployment ensures that DHCP services are available to remote telephony devices even during WAN failures. Example 3-1 lists the basic Cisco IOS DHCP server configuration commands.

Example 3-1 Cisco IOS DHCP Server Configuration Commands

```
! Activate DHCP Service on the IOS Device

service dhcp

! Specify any IP Address or IP Address Range to be excluded from the DHCP pool

ip dhcp excluded-address <ip-address>|<ip-address-low> <ip-address-high>

! Specify the name of this specific DHCP pool, the subnet and mask for this
! pool, the default gateway and up to four TFTP

ip dhcp pool <dhcp-pool name>
   network <ip-subnet> <mask>
   default-router <default-gateway-ip>
   option 150 ip <tftp-server-ip-1> ...

! Note: IP phones use only the first two addresses supplied in the option 150
! field even if more than two are configured.
```

Unified CM DHCP Sever (Standalone versus Co-Resident DHCP)

Typically DHCP servers are dedicated machine(s) in most network infrastructures, and they run in conjunction with the DNS and/or the Windows Internet Naming Service (WINS) services used by that network. In some instances, given a small Unified CM deployment with no more than 1000 devices registering to the cluster, you may run the DHCP server on a Unified CM server to support those devices. However, to avoid possible resource contention such as CPU contention with other critical services running on Unified CM, Cisco recommends moving the DHCP Server functionality to a dedicated server. If more than 1000 devices are registered to the cluster, DHCP must *not* be run on a Unified CM server but instead must be run on a dedicated or standalone server(s).

Note The term *co-resident* refers to two or more services or applications running on the same server or virtual machine.

Trivial File Transfer Protocol (TFTP)

Within a Cisco Unified CM system, endpoints such as IP phones rely on a TFTP-based process to acquire configuration files, software images, and other endpoint-specific information. The Cisco TFTP service is a file serving system that can run on one or more Unified CM servers. It builds configuration files and serves firmware files, ringer files, device configuration files, and so forth, to endpoints.

The TFTP file systems can hold several file types, such as the following:

- Phone configuration files
- Phone firmware files
- Certificate Trust List (CTL) files
- Identity Trust List (ITL) files
- Tone localization files
- User interface (UI) localization and dictionary files
- Ringer files
- Softkey files
- Dial plan files for SIP phones

The TFTP server manages and serves two types of files, those that are not modifiable (for example, firmware files for phones) and those that can be modified (for example, configuration files).

A typical configuration file contains a prioritized list of Unified CMs for a device (for example, an SCCP or SIP phone), the TCP ports on which the device connects to those Unified CMs, and an executable load identifier. Configuration files for selected devices contain locale information and URLs for the messages, directories, services, and information buttons on the phone.

When a device's configuration changes, the TFTP server rebuilds the configuration files by pulling the relevant information from the Unified CM database. The new file(s) is then downloaded to the phone once the phone has been reset. As an example, if a single phone's configuration file is modified (for example, during Extension Mobility login or logout), only that file is rebuilt and downloaded to the phone. However, if the configuration details of a device pool are changed (for example, if the primary Unified CM server is changed), then all devices in that device pool need to have their configuration files rebuilt and downloaded. For device pools that contain large numbers of devices, this file rebuilding process can impact server performance.

Note The TFTP server can perform a local database read from the database on its co-resident subscriber server. Local database read not only provides benefits such as the preservation of user-facing features when the publisher in unavailable, but also allows multiple TFTP servers to be distributed by means of clustering over the WAN. (The same latency rules for clustering over the WAN apply to TFTP servers as apply to servers with registered phones.) This configuration brings the TFTP service closer to the endpoints, thus reducing latency and ensuring failure isolation between the sites.

When a device requests a configuration file from the TFTP server, the TFTP server searches for the configuration file in its internal caches, the disk, and then remote Cisco TFTP servers (if specified). If the TFTP server finds the configuration file, it sends it to the device. If the configuration file provides Unified CM names, the device resolves the name by using DNS and opens a connection to the Unified CM. If the device does not receive an IP address or name, it uses the TFTP server name or IP address to attempt a registration connection. If the TFTP server cannot find the configuration file, it sends a "file not found" message to the device.

A device that requests a configuration file while the TFTP server is processing the maximum number of requests, will receive a message from the TFTP server that causes the device to request the configuration file later. The Maximum Serving Count service parameter, which can be configured, specifies the maximum number of requests that can be concurrently handled by the TFTP server. (Default value = 2,500 requests.) Use the default value if the TFTP service is run along with other Cisco CallManager services on the same server. For a dedicated TFTP server, use the following suggested values for the Maximum Serving Count: 2,500 for a single-processor system or 3,000 for a dual-processor system.

The Cisco Unified IP Phones 8900 Series and 9900 Series request their TFTP configuration files over the HTTP protocol (port 6970), which is much faster than TFTP.

An Example of TFTP in Operation

Every time an endpoint reboots, the endpoint will request a configuration file (via TFTP) whose name is based on the requesting endpoint's MAC address. (For a Cisco Unified IP Phone 7961 with MAC address ABCDEF123456, the file name would be SEPABCDEF123456.cnf.xml.) The received configuration file includes the version of software that the phone must run and a list of Cisco Unified CM servers with which the phone should register. The endpoint might also download, via TFTP, ringer files, softkey templates, and other miscellaneous files to acquire the necessary configuration information before becoming operational.

If the configuration file includes software file(s) version numbers that are different than those the phone is currently using, the phone will also download the new software file(s) from the TFTP server to upgrade itself. The number of files an endpoint must download to upgrade its software varies based on the type of endpoint and the differences between the phone's current software and the new software.

TFTP File Transfer Times

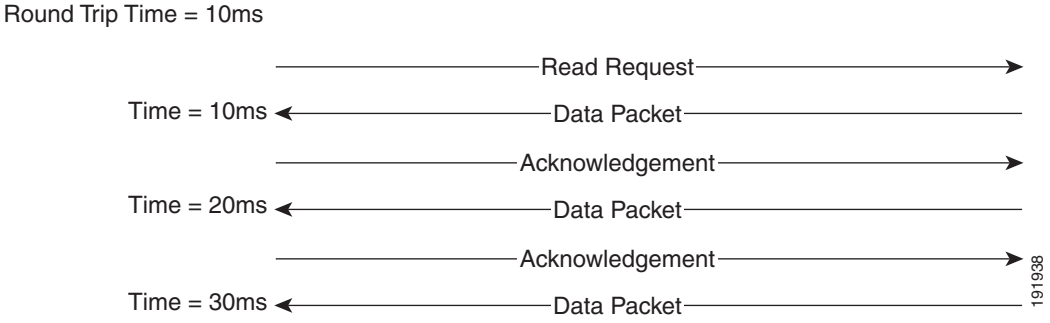
Each time an endpoint requests a file, there is a new TFTP transfer session. For centralized call processing deployments, the time to complete each of these transfers will affect the time it takes for an endpoint to start and become operational as well as the time it takes for an endpoint to upgrade during a scheduled maintenance. While TFTP transfer times are not the only factor that can affect these end states, they are a significant component.

The time to complete each file transfer via TFTP is predictable as a function of the file size, the percentage of TFTP packets that must be retransmitted, and the network latency or round-trip time.

At first glance, network bandwidth might seem to be missing from the previous statement, but it is actually included via the percentage of TFTP packets that must be retransmitted. This is because, if there is not enough network bandwidth to support the file transfer(s), then packets will be dropped by the network interface queuing algorithms and will have to be retransmitted.

TFTP operates on top of the User Datagram Protocol (UDP). Unlike Transmission Control Protocol (TCP), UDP is not a reliable protocol, which means that UDP does not inherently have the ability to detect packet loss. Obviously, detecting packet loss in a file transfer is important, so RFC 1350 defines TFTP as a lock-step protocol. In other words, a TFTP sender will send one packet and wait for a response before sending the next packet (see Figure 3-9).

Figure 3-9 Example of TFTP Packet Transmission Sequence

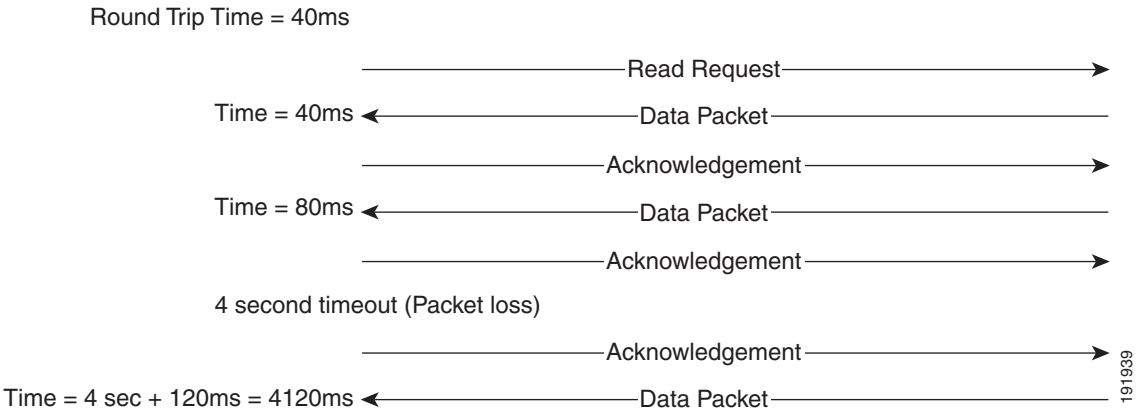


If a response is not received in the timeout period (4 seconds by default), the sender will resend the data packet or acknowledgment. When a packet has been sent five times without a response, the TFTP session fails. Because the timeout period is always the same and not adaptive like a TCP timeout, packet loss can significantly increase the amount of time a transfer session takes to complete.

Because the delay between each data packet is, at a minimum, equal to the network round-trip time, network latency also is a factor in the maximum throughput that a TFTP session can achieve.

In Figure 3-10, the round-trip time has been increased to 40 ms and one packet has been lost in transit. While the error rate is high at 12%, it is easy to see the effect of latency and packet loss on TFTP because the time to complete the session increased from 30 ms (in Figure 3-9) to 4160 ms (in Figure 3-10).

Figure 3-10 Effect of Packet Loss on TFTP Session Completion Time



Use the following formula to calculate how long a TFTP file transfer will take to complete:

FileTransferTime = FileSize * [(RTT + ERR * Timeout) / 512000]

Where:

FileTransferTime is in seconds.

FileSize is in bytes.

RTT is the round-trip time in milliseconds.

ERR is the error rate, or percentage of packets that are lost.

Timeout is in milliseconds.

512000 = (TFTP packet size) * (1000 millisecond per seconds) =
(512 bytes) * (1000 millisecond per seconds)

Cisco Unified IP Phone Firmware Releases 7.*x* have a 10-minute timeout when downloading new files. If the transfer is not completed within this time, the phone will discard the download even if the transfer completes successfully later. If you experience this problem, Cisco recommends that you use a local TFTP server to upgrade phones to the 8.*x* firmware releases, which have a timeout value of 61 minutes.

Because network latency and packet loss have such an effect on TFTP transfer times, a local TFTP Server can be advantageous. This local TFTP server may be a Unified CM subscriber in a deployment with cluster over the WAN or an alternative local TFTP "Load Server" running on a Cisco Integrated Services Router (ISR), for example. Newer endpoints (which have larger firmware files) can be configured with a Load Server address, which allows the endpoint to download the relatively small configuration files from the central TFTP server but use a local TFTP Server (which is not part of the Unified CM cluster) to download the larger software files. For details on which Cisco Unified IP Phones support an alternative local TFTP Load Server, refer to the product documentation for your particular phone models (available at https://www.cisco.com).

Note The exact process each phone goes through on startup and the size of the files downloaded will depend on the phone model, the signaling type configured for the phone (SCCP, MGCP, or SIP) and the previous state of the phone. While there are differences in which files are requested, the general process each phone follows is the same, and in all cases the TFTP server is used to request and deliver the appropriate files. The general recommendations for TFTP server deployment do not change based on the protocol and/or phone models deployed.

TFTP Server Redundancy

Option 150 allows up to two IP addresses to be returned to phones as part of the DHCP scope. The phone tries the first address in the list, and it tries the subsequent address only if it cannot establish communications with the first TFTP server. This address list provides a redundancy mechanism that enables phones to obtain TFTP services from another server even if their primary TFTP server has failed.

TFTP Load Sharing

Cisco recommends that you grant different ordered lists of TFTP servers to different subnets to allow for load balancing. For example:

- In subnet 10.1.1.0/24: Option 150: TFTP1_Primary, TFTP1_Secondary
- In subnet 10.1.2.0/24: Option 150: TFTP1_Secondary, TFTP1_Primary

Under normal operations, a phone in subnet 10.1.1.0/24 will request TFTP services from TFTP1_Primary, while a phone in subnet 10.1.2.0/24 will request TFTP services from TFTP1_Secondary. If TFTP1_Primary fails, then phones from both subnets will request TFTP services from TFTP1_Secondary.

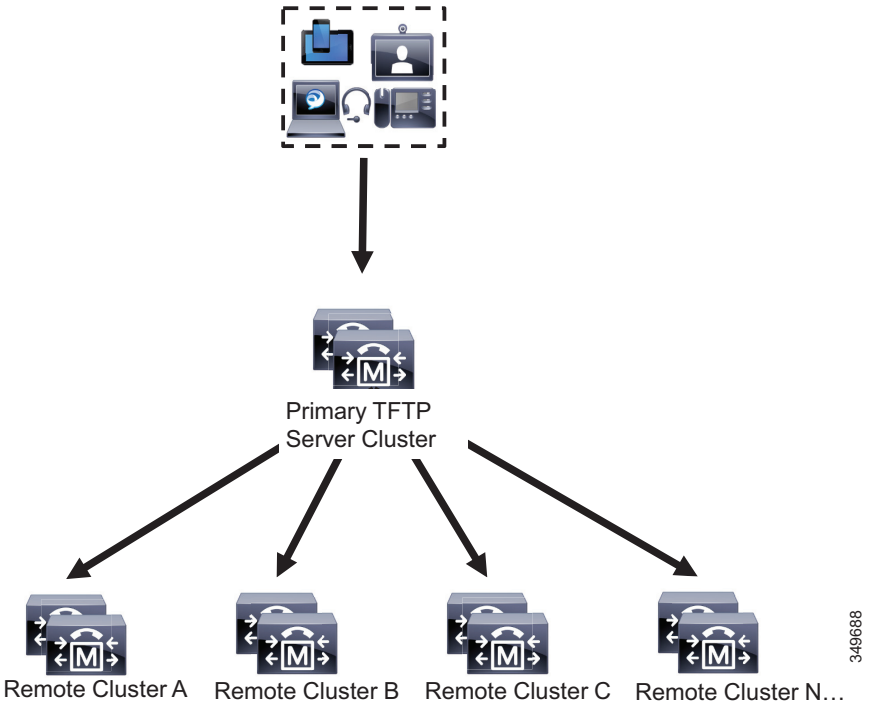
Load balancing avoids having a single TFTP server hot-spot, where all phones from multiple clusters rely on the same server for service. TFTP load balancing is especially important when phone software loads are transferred, such as during a Unified CM upgrade, because more files of larger size are being transferred, thus imposing a bigger load on the TFTP server.

Proxy TFTP

In multi-cluster systems, the proxy TFTP service is able provide TFTP files from multiple clusters via a single primary TFTP server. The proxy TFTP can serve as a single TFTP reference for scenarios where a single subnet or VLAN contains phones from multiple clusters or in any scenario where multiple clusters share the same DHCP TFTP option (150).

The Proxy TFTP service functions as a single-level hierarchy, as illustrated in Figure 3-11. More complicated multi-level hierarchies are not supported.

Figure 3-11 Proxy TFTP Single-Level Hierarchy



In Figure 3-11 a group of devices contacts the Primary TFTP server for their configuration files. When it receives a request for TFTP from a device, the primary TFTP looks into its own local cache for the configuration file as well as any other remotely configured clusters such as Remote Cluster A, B, C, or N (any other remote clusters configured).

It is possible to configure any number of remote clusters on the primary TFTP server; however, each remote cluster may contain only up to 3 TFTP IP addresses. The recommended design for redundancy is 2 TFTP servers per cluster, and thus 2 IP addresses per remote cluster on the Primary TFTP server for redundancy.

Network Time Protocol (NTP)

NTP allows network devices to synchronize their clocks to a network time server or network-capable clock. NTP is critical for ensuring that all devices in a network have the same time. When troubleshooting or managing a telephony network, it is crucial to synchronize the time stamps within all error and security logs, traces, and system reports on devices throughout the network. This synchronization enables administrators to recreate network activities and behaviors based on a common timeline. Billing records and call detail records (CDRs) also require accurate synchronized time.

Unified CM NTP Time Synchronization

Time synchronization is especially critical on Unified CM servers. In addition to ensuring that CDR records are accurate and that log files are synchronized, having an accurate time source is necessary for any future IPSec features to be enabled within the cluster and for communications with any external entity.

Unified CM automatically synchronizes the NTP time of all subscribers in the cluster to the publisher. During installation, each subscriber is automatically configured to point to an NTP server running on the publisher. The publisher considers itself to be a master server and provides time for the cluster based on its internal hardware clock unless it is configured to synchronize from an external server. Cisco highly recommends configuring the publisher to point to a Stratum-1, Stratum-2, or Stratum-3 NTP server to ensure that the cluster time is synchronized with an external time source.

Using Windows Time Services as an NTP server is not recommended or supported because Windows Time Services often use Simple Network Time Protocol (SNTP), and Cisco Unified CM cannot successfully synchronize with SNTP.

The external NTP server specified for the primary node should be NTP v4 (version 4) to avoid potential compatibility, accuracy, and network jitter problems. External NTP servers *must* be NTP v4 if IPv6 addressing is used.

Cisco IOS and CatOS NTP Time Synchronization

Time synchronization is also important for other devices within the network. Cisco IOS routers and Catalyst switches should be configured to synchronize their time with the rest of the network devices via NTP. This is critical for ensuring that debug, syslog, and console log messages are time-stamped appropriately. Troubleshooting telephony network issues is simplified when a clear timeline can be drawn for events that occur on devices throughout the network.

WAN Infrastructure

Proper WAN infrastructure design is also extremely important for normal Unified Communications operation on a converged network. Proper infrastructure design requires following basic configuration and design best practices for deploying a WAN that is as highly available as possible and that provides guaranteed throughput. Furthermore, proper WAN infrastructure design requires deploying end-to-end QoS on all WAN links. The following sections discuss these requirements:

- WAN Design and Configuration, page 3-34
- WAN Quality of Service (QoS), page 3-37
- Bandwidth Provisioning, page 3-52

For more information on bandwidth management, see the chapter on Bandwidth Management, page 13-1.

WAN Design and Configuration

Properly designing a WAN requires building fault-tolerant network links and planning for the possibility that these links might become unavailable. By carefully choosing WAN topologies, provisioning the required bandwidth, and approaching the WAN infrastructure as another layer in the network topology, you can build a fault-tolerant and redundant network. The following sections examine the required infrastructure layers and network services:

- Deployment Considerations, page 3-34
- Guaranteed Bandwidth, page 3-35
- Best-Effort Bandwidth, page 3-36

Deployment Considerations

WAN deployments for voice and video networks may use a hub-and-spoke, fully meshed, or partially meshed topology. A hub-and-spoke topology consists of a central hub site and multiple remote spoke sites connected into the central hub site. In this scenario, each remote or spoke site is one WAN-link hop away from the central or hub site and two WAN-link hops away from all other spoke sites. A meshed topology may contain multiple WAN links and any number of hops between the sites. In this scenario there may be many different paths to the same site or there may be different links used for communication with some sites compared to other sites. The simplest example is three sites, each with a WAN link to the other two sites, forming a triangle. In that case there are two potential paths between each site to each other site.

For more information about centralized and distributed multisite deployment models as well as Multiprotocol Label Switching (MPLS) implications for these deployment models, see the chapter on Collaboration Deployment Models, page 10-1.

WAN links should, when possible, be made redundant to provide higher levels of fault tolerance. Redundant WAN links provided by different service providers or located in different physical ingress/egress points within the network can ensure backup bandwidth and connectivity in the event that a single link fails. In non-failure scenarios, these redundant links may be used to provide additional bandwidth and offer load balancing of traffic on a per-flow basis over multiple paths and equipment within the WAN.

Voice, video, and data should remain converged at the WAN, just as they are converged at the LAN. QoS provisioning and queuing mechanisms are typically available in a WAN environment to ensure that voice, video, and data can interoperate on the same WAN links. Attempts to separate and forward voice, video, and data over different links can be problematic in many instances because the failure of one link typically forces all traffic over a single link, thus diminishing throughput for each type of traffic and in most cases reducing the quality of voice. Furthermore, maintaining separate network links or devices makes troubleshooting and management difficult at best.

Because of the potential for WAN links to fail or to become oversubscribed, Cisco recommends deploying non-centralized resources as appropriate at sites on the other side of the WAN. Specifically, media resources, DHCP servers, voice gateways, and call processing applications such as Survivable Remote Site Telephony (SRST) and Cisco Unified Communications Manager Express (Unified CME) should be deployed at non-central sites when and if appropriate, depending on the site size and how critical these functions are to that site. Keep in mind that de-centralizing voice applications and devices can increase the complexity of network deployments, the complexity of managing these resources throughout the enterprise, and the overall cost of a the network solution; however, these factors can be mitigated by the fact that the resources will be available during a WAN link failure.

When deploying voice in a WAN environment, it is possible to reduce bandwidth consumption by using the lower-bandwidth G.729 codec for any voice calls that will traverse WAN links because this practice will provide bandwidth savings on these lower-speed links. Furthermore, media resources such as MoH can also be configured to use multicast transport mechanism when possible because this practice will provide additional bandwidth savings.

Delay in IP Voice Networks

Recommendation G.114 of the International Telecommunication Union (ITU) states that the one-way delay in a voice network should be less than or equal to 150 milliseconds. It is important to keep this in mind when implementing low-speed WAN links within a network. Topologies, technologies, and physical distance should be considered for WAN links so that one-way delay is kept at or below this 150-millisecond recommendation. Implementing a VoIP network where the one-way delay exceeds 150 milliseconds introduces issues not only with the quality of the voice call but also with call setup and media cut-through times because several call signaling messages need to be exchanged between each device and the call processing application in order to establish the call.

Guaranteed Bandwidth

Because voice is typically deemed a critical network application, it is imperative that bearer and signaling voice traffic always reaches its destination. For this reason, it is important to choose a WAN topology and link type that can provide guaranteed dedicated bandwidth. The following WAN link technologies can provide guaranteed dedicated bandwidth:

- Leased Lines
- Frame Relay
- Asynchronous Transfer Mode (ATM)
- ATM/Frame-Relay Service Interworking
- Multiprotocol Label Switching (MPLS)
- Cisco Voice and Video Enabled IP Security VPN (IPSec V3PN)

These link technologies, when deployed in a dedicated fashion or when deployed in a private network, can provide guaranteed traffic throughput. All of these WAN link technologies can be provisioned at specific speeds or bandwidth sizes. In addition, these link technologies have built-in mechanisms that help guarantee throughput of network traffic even at low link speeds. Features such as traffic shaping, fragmentation and packet interleaving, and committed information rates (CIR) can help ensure that packets are not dropped in the WAN, that all packets are given access at regular intervals to the WAN link, and that enough bandwidth is available for all network traffic attempting to traverse these links.

Dynamic Multipoint VPN (DMVPN)

Spoke-to-spoke DMVPN networks can provide benefits for Cisco Unified Communications compared with hub-and-spoke topologies. Spoke-to-spoke tunnels can provide a reduction in end-to-end latency by reducing the number of WAN hops and decryption/encryption stages. In addition, DMVPN offers a simplified means of configuring the equivalent of a full mesh of point-to-point tunnels without the associated administrative and operational overhead. The use of spoke-to-spoke tunnels also reduces traffic at the hub, thus providing bandwidth and router processing capacity savings. Spoke-to-spoke DMVPN networks, however, are sensitive to the delay variation (jitter) caused during the transition of RTP packets routing from the spoke-hub-spoke path to the spoke-to-spoke path. This variation in delay during the DMVPN path transition occurs very early in the call and is generally unnoticeable, although a single momentary audio distortion might be heard if the latency difference is above 100 ms.

For information on the deployment of multisite DMVPN WANs with centralized call processing, refer to the *Cisco Unified Communications Voice over Spoke-to-Spoke DMVPN Test Results and Recommendations*, available at https://www.cisco.com/go/designzone.

Best-Effort Bandwidth

There are some WAN topologies that are unable to provide guaranteed dedicated bandwidth to ensure that network traffic will reach its destination, even when that traffic is critical. These topologies are extremely problematic for voice traffic, not only because they provide no mechanisms to provision guaranteed network throughput, but also because they provide no traffic shaping, packet fragmentation and interleaving, queuing mechanisms, or end-to-end QoS to ensure that critical traffic such as voice will be given preferential treatment.

The following WAN network topologies and link types are examples of this kind of best-effort bandwidth technology:

- The Internet
- DSL
- Cable
- Satellite
- Wireless

In most cases, none of these link types can provide the guaranteed network connectivity and bandwidth required for critical voice and voice applications. However, these technologies might be suitable for personal or telecommuter-type network deployments. At times, these topologies can provide highly available network connectivity and adequate network throughput; but at other times, these topologies can become unavailable for extended periods of time, can be throttled to speeds that render network throughput unacceptable for real-time applications such as voice, or can cause extensive packet losses and require repeated retransmissions. In other words, these links and topologies are unable to provide guaranteed bandwidth, and when traffic is sent on these links, it is sent best-effort with no guarantee that it will reach its destination. For this reason, Cisco recommends that you do *not* use best-effort WAN topologies for voice-enabled networks that require enterprise-class voice services and quality.

Note There are some new QoS mechanisms for DSL and cable technologies that can provide guaranteed bandwidth; however, these mechanisms are not typically deployed by many service providers. For any service that offers QoS guarantees over networks that are typically based on best-effort, it is important to review and understand the bandwidth and QoS guarantees offered in the service provider's service level agreement (SLA).

Note Upstream and downstream QoS mechanisms are now supported for wireless networks. For more information on QoS for Voice over Wireless LANs, refer to the *Voice over Wireless LAN Design Guide*, available at https://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns818/landing_wireless_uc.html.

WAN Quality of Service (QoS)

The case for Quality of Service over the enterprise WAN and VPN is largely self-evident, as these links are often orders of magnitude slower than the (Gigabit or Ten-Gigabit Ethernet) campus or branch LAN links to which they connect. As such, these WAN and VPN edges usually represent the greatest bottlenecks in the network and therefore require the most attention to QoS design.

Two key strategic QoS design principles are highly applicable to WAN/VPN QoS design:

- Enable queuing policies at every node where the potential for congestion exists, which generally equates to attaching a comprehensive queuing policy to every WAN/VPN edge.
- Protect the control plane and data plane by enabling control plane policing (on platforms supporting this feature) as well as data plane policing (scavenger class QoS) to mitigate and constrain network attacks.

To this end, this design section provides best-practice recommendations for enabling QoS over the wide area network. However, it is important to note that the recommendations in this section are not autonomous, but rather, they depend on the campus QoS design recommendations presented in the section on LAN Quality of Service (QoS), page 3-14, having already been implemented. Traffic traversing the WAN can thus be assumed to be correctly classified and marked with Layer 3 DSCP (as well as policed at the access-edge, as necessary).

Furthermore, this design section covers fundamental considerations relating to wide area networks. Before strategic QoS designs for the WAN can be derived, a few WAN-specific considerations need to be taken into account, as are discussed below. Further information on bandwidth management in a Collaboration solution can be found in the chapter on Bandwidth Management, page 13-1.

WAN QoS Design Considerations

Several considerations factor into WAN and VPN QoS designs, including:

- WAN Aggregation Router Platforms, page 3-37
- Hardware versus Software QoS, page 3-38
- Latency and Jitter, page 3-38
- Tx-Ring, page 3-40
- Class-Based Weighted-Fair Queuing, page 3-41
- Low-Latency Queuing, page 3-43
- Weighted-Random Early Detect, page 3-44

Each of these WAN QoS design considerations is discussed in the following sections.

WAN Aggregation Router Platforms

Extending an enterprise campus network over a wide area to interconnect with other campus and/or branch networks usually requires two types of routers to be deployed: WAN aggregation routers and branch routers. WAN aggregation routers serve to connect large campus networks to the WAN/VPN, whereas branch routers serve to connect smaller branch LANs to the WAN/VPN.

Hardware versus Software QoS

Unlike Cisco Catalyst switches utilized within the campus, which perform QoS exclusively in hardware, Cisco routers perform QoS operations in Cisco IOS software, although some platforms (such as the Cisco Catalyst 6500 Series, 7600 Series, and Cisco ASRs) perform QoS in a hybrid mix of software and hardware.

Performing QoS in Cisco IOS software allows for several advantages, including:

- Cross-platform consistency in QoS features

  For example, rather than having hardware-specific queuing structures on a per-platform or per-line-card basis (as is the case for Cisco Catalyst switches), standard software queuing features such as Low-Latency Queuing (LLQ) and Class-Based Weighted-Fair Queuing (CBWFQ) can be utilized across WAN and branch router platforms.

- Consistent QoS configuration syntax

  The configuration syntax for Cisco IOS QoS, namely the Modular QoS Command Line Interface (MQC) syntax is (with very few exceptions) identical across these WAN and branch router platforms.

- Richer QoS features

  Many Cisco IOS QoS features such as Network Based Application Recognition (NBAR) and Hierarchical QoS (HQoS) are not available on most Catalyst hardware platforms.

Latency and Jitter

Some real-time applications have fixed latency budgets; for example, the ITU G.114 specification sets the target for one-way latency for real-time voice/video conversations to be 150 ms. In order to meet such targets, it is important for administrators to understand the components of network latency so they know which factors they can and cannot control with the network and QoS design. Network latency can be divided into fixed and variable components:

- Serialization (fixed)
- Propagation (fixed)
- Queuing (variable)

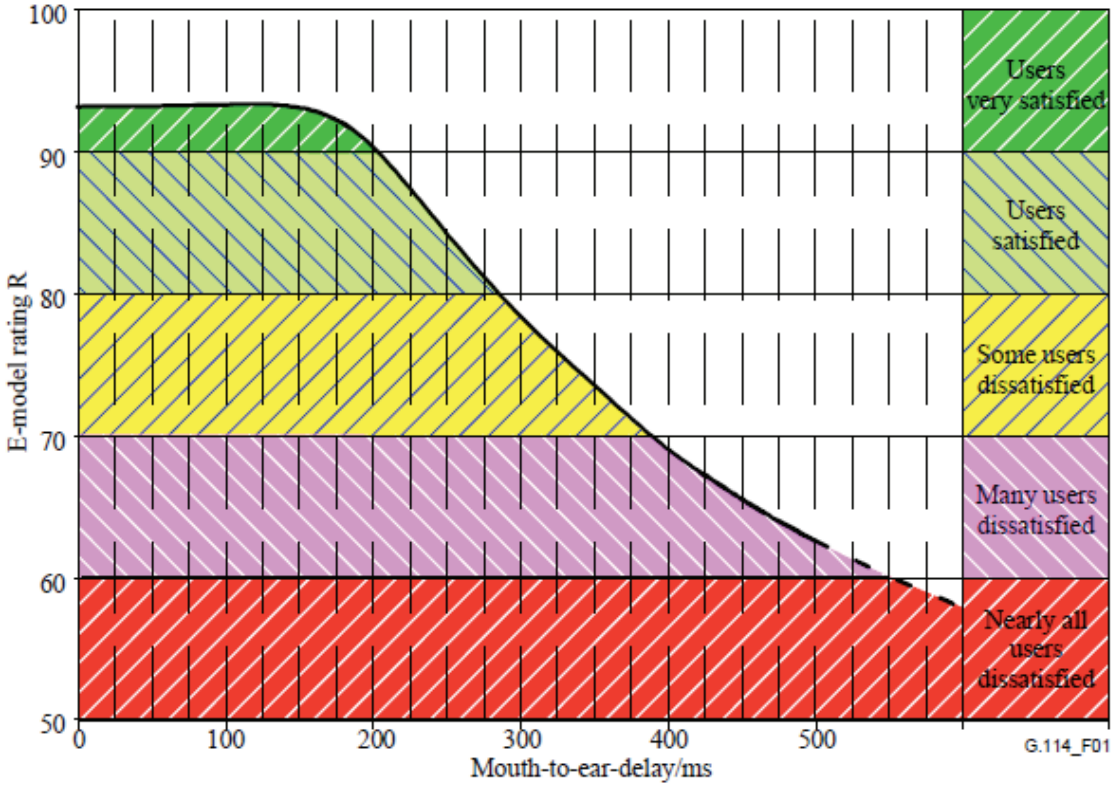
Serialization refers to the time it takes to convert a Layer 2 frame into Layer 1 electrical or optical pulses onto the transmission media. Therefore, serialization delay is fixed and is a function of the line rate (that is, the clock speed of the link). For example, a (1.544 Mbps) T1 circuit would require about 8 ms to serialize a 1,500 byte Ethernet frame onto the wire, whereas a (9.953 Gbps) OC-192/STM-64 circuit would require just 1.2 microseconds to serialize the same frame.

Usually, the most significant network factor in meeting the latency targets for over the WAN is propagation delay, which can account for over 95% of the network latency time budget. Propagation delay is also a fixed component and is a function of the physical distance that the signals have to travel between the originating endpoint and the receiving endpoint. The gating factor for propagation delay is the speed of light, which is 300,000 km/s or 186,000 miles per second in a vacuum. However, the speed of light in an optical fiber is about one third the speed of light in a vacuum. Thus, the propagation delay for most fiber circuits is approximately 6.3 microseconds per km or 8.2 microseconds per mile.

Another point to keep in mind when calculating propagation delay is that optical fibers are not always physically placed over the shortest path between two geographic points, especially over transoceanic links. Due to installation convenience, circuits may be hundreds or even thousands of miles longer than theoretically necessary.

Nonetheless, the G.114 real-time communications network latency budget of 150 ms allows for nearly 24,000 km or 15,000 miles worth of propagation delay (which is approximately 60% of the earth's circumference). The theoretical worst-case scenario (exactly half of the earth's circumference) would require only 126 ms of latency. Therefore, this latency target is usually achievable for virtually any two locations (via a terrestrial path), given relatively direct transmission paths; however, in some scenarios meeting this latency target might simply not be possible due to the distances involved and the relative directness of their respective transmission paths. In such scenarios, if the G.114 150 ms one-way latency target cannot be met due to the distances involved, administrators should be aware that both the ITU and Cisco Technical Marketing have shown that real-time communication quality does not begin to degrade significantly until one-way latency exceeds 200 ms, as is illustrated in the ITU G.114 graph of real-time speech quality versus absolute delay, which is reproduced in Figure 3-12.

Figure 3-12 ITU G.114 Graph of Real-time Speech Quality versus Latency



Source: *ITU-T Recommendation G.114* (05/2003), available at http://www.itu.int/rec/T-REC-G.114-200305-I/en

348825

Note This discussion so far has focused on WAN circuits over terrestrial paths. For satellite circuits, the expected latency can be in the range of 250 to 900 ms. For example, signals being relayed via geostationary satellites will need to be sent to an altitude of 35,786 km (22,236 miles) above sea level (from the equator) out into space and then back to Earth again. There is nothing an administrator can do

to decrease latency in such scenarios because they can do nothing about increasing the speed of light or radio waves. All that can be done to address the effect of latency in these scenarios is to educate the user-base so that realistic performance expectations are set.

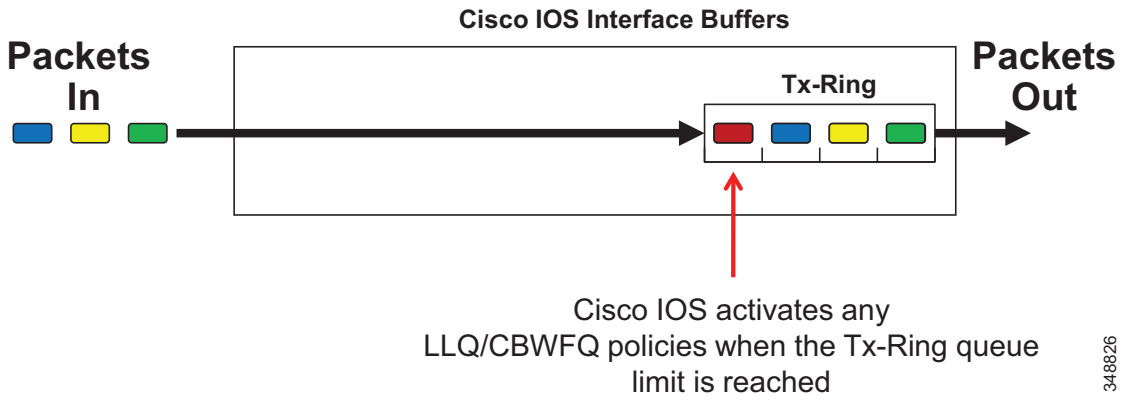
The final network latency component to be considered is queuing delay, which is variable (variable delay is also known as jitter). Queuing delay is a function of whether a network node is congested and, if so, what scheduling policies have been applied to resolve congestion events. Real-time applications are often more sensitive to jitter than latency, because packets need to be received in de-jitter buffers prior to being played out. If a packet is not received within the time allowed by the de-jitter buffer, it is essentially lost and can affect the overall voice or video call quality.

Given that the majority of factors contributing to network latency are fixed, careful attention has to be given to queuing delay, since this is the only latency factor that is directly under the network administrator's control via queuing policies. Therefore, a close examination of the Cisco IOS queuing system, including the Tx-Ring and LLQ/CBWFQ operation, will assist network administrators to optimize these critical policies.

Tx-Ring

The Tx-Ring is the final Cisco IOS output buffer for a WAN interface (a relatively small FIFO queue), and it maximizes physical link bandwidth utilization by matching the outbound packet rate on the router with the physical interface rate. The Tx-Ring is illustrated in Figure 3-13.

Figure 3-13 Cisco IOS Tx-Ring Operation



The Tx-Ring also serves to indicate interface congestion to the Cisco IOS software. Prior to interface congestion, packets are sent on a FIFO basis to the interface via the Tx-Ring. However, when the Tx-Ring fills to its queue limit, then it signals to the Cisco IOS software to engage any LLQ or CBWFQ policies that have been attached to the interface. Subsequent packets are then queued within Cisco IOS according to these LLQ and CBWFQ policies, dequeued into the Tx-Ring, and then sent out the interface in a FIFO manner.

The Tx-Ring can be configured on certain platforms with the **tx-ring-limit** interface configuration command. The default value of the Tx-Ring varies according to platform and link type and speed. For further details, refer to *Understanding and Tuning the tx-ring-limit Value*, available at

https://www.cisco.com/c/en/us/support/docs/asynchronous-transfer-mode-atm/ip-to-atm-class-of-service/6142-txringlimit-6142.html

Changing the Tx-Ring Default Setting

During Cisco Technical Marketing design validation, it was observed that the default Tx-Ring limit on some interfaces caused somewhat higher jitter values to some real-time application classes, particularly HD video-based real-time applications such as Cisco TelePresence traffic. The reason for this is the bursty nature of HD video traffic. For example, consider a fully-congested T3 WAN link (using a Cisco PA-T3+ port adapter interface) with active LLQ and CBWFQ policies. The default Tx-Ring depth in this case is 64 packets. Even if TelePresence traffic is prioritized via an LLQ, if there are no TelePresence packets to send, the FIFO Tx-Ring is filled with other traffic to a default depth of 64 packets. When a new TelePresence packet arrives, even if it gets priority treatment from the Layer 3 LLQ/CBWFQ queuing system, the packets are dequeued into the FIFO Tx-Ring when space is available. However, with the default settings, there can be as many as 63 packets in the Tx-Ring in front of that TelePresence packet. In such a worst-case scenario it could take as long as 17 ms to transmit these non-real-time packets out of this (45 Mbps) T3 interface. This 17 ms of instantaneous and variable delay (jitter) can affect the video quality for TelePresence to the point of being visually apparent to the end user. However, lowering the value of the Tx-Ring on this link will force in the Cisco IOS software engaging congestion management policies sooner and more often, resulting in lower overall jitter values for real-time applications such as TelePresence.

On the other hand, setting the value of the Tx-Ring too low might result in significantly higher CPU utilization rates because the processor is continually being interrupted to engage queuing policies, even when congestion rates are just momentary bursts and not sustained rates. Thus, when tuning the Tx-Ring, a trade-off setting is required so that jitter is minimized, but not at the expense of excessive CPU utilization rates.

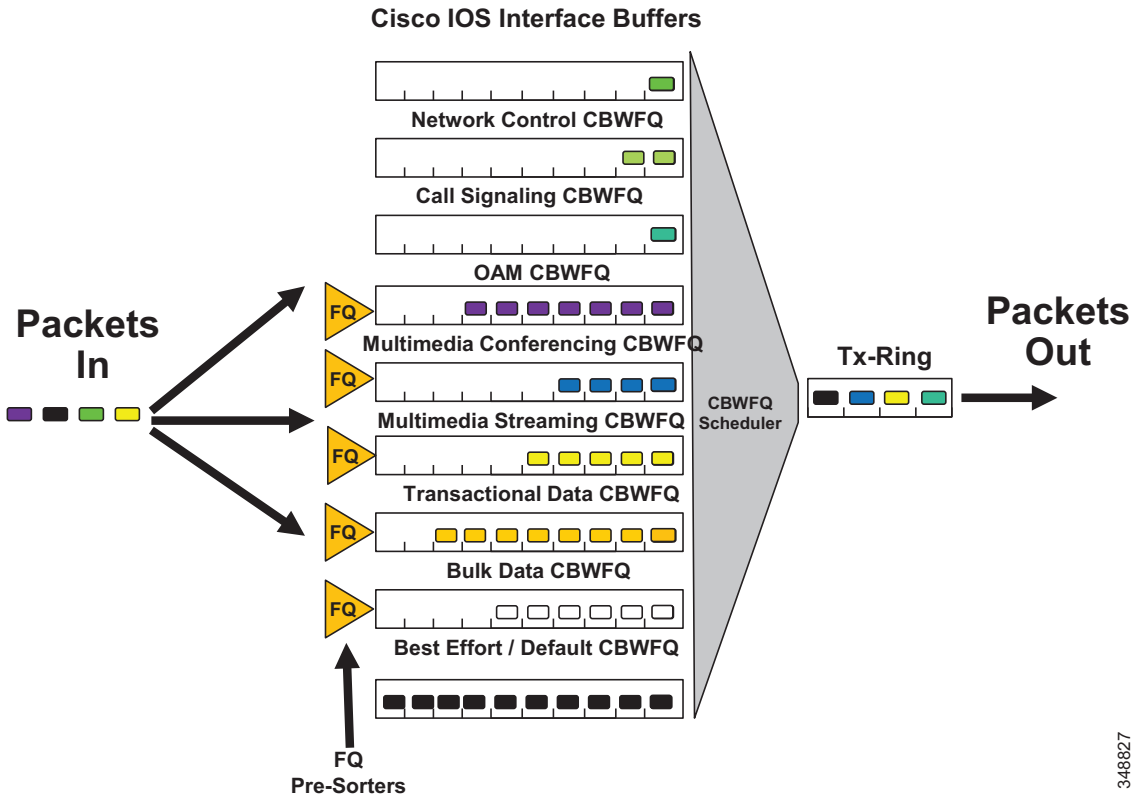
Therefore, explicit attention needs be given to link types and speeds when the Tx-Ring is tuned away from default values.

Class-Based Weighted-Fair Queuing

Class-Based Weighted-Fair Queuing (CBWFQ) is a Cisco IOS queuing algorithm that combines the ability to guarantee bandwidth with the ability to dynamically ensure fairness to other flows within a class of traffic.

The Cisco IOS software engages CBWFQ policies (provided they have been attached to an interface) only if the Tx-Ring for the interface is full, which occurs only in the event of congestion. Once congestion has thus been signaled to the software, each CBWFQ class is assigned its own queue. CBWFQ queues may also have a fair-queuing pre-sorter applied to them, so that multiple flows contending for a single queue are managed fairly. Additionally, each CBWFQ queue is serviced in a Weighted-Round-Robin (WRR) fashion based on the bandwidth assigned to each class. The CBWFQ scheduler then forwards packets to the Tx-Ring. The operation of CBWFQ is illustrated in Figure 3-14.

Figure 3-14 Cisco IOS CBWFQ Operation



Each CBWFQ class is guaranteed bandwidth via a **bandwidth** policy-map class configuration statement. CBWFQ derives the weight for packets belonging to a class from the bandwidth allocated to the class. CBWFQ then uses the weight to ensure that the queue for the class is serviced fairly, via WRR scheduling.

An important point regarding bandwidth assigned to a given CBWFQ class is that the bandwidth allocated is not a static bandwidth reservation, but rather represents a minimum bandwidth guarantee to the class, provided there are packets offered to the class. If there are no packets offered to the class, then the scheduler services the next queue and can dynamically redistribute unused bandwidth allocations to other queues as necessary.

Additionally, a fair-queuing pre-sorter may be applied to specific CBWFQ queues with the **fair-queue** policy-map class configuration command. It should be noted that this command enables a flow-based fair-queuing pre-sorter, and not a weighted fair-queuing pre-sorter, as the name for this feature implies (and as such, the fair-queuing pre-sorter does not take into account the IP Precedence values of any packets offered to a given class). For example, if a CBWFQ class was assigned 1 Mbps of bandwidth and there were 4 competing traffic flows contending for this class, a fair-queuing pre-sorter would ensure that each flow receives (1 / (total-number-of-flows)) of bandwidth, or in this example (1/4 of 1 Mpbs) 250 kbps of bandwidth.

Note Prior to Cisco IOS Release 12.4(20)T, a fair-queue pre-sorter could be applied only to class-default; however, subsequent Cisco IOS releases include the support of the Hierarchical Queuing Framework (HQF) which, among many other QoS feature enhancements, allows for a fair-queue pre-sorter to be applied to any CBWFQ class. HQF details are documented at https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/qos_hrhqf/configuration/15-mt/qos-hrhqf-15-mt-book.html.

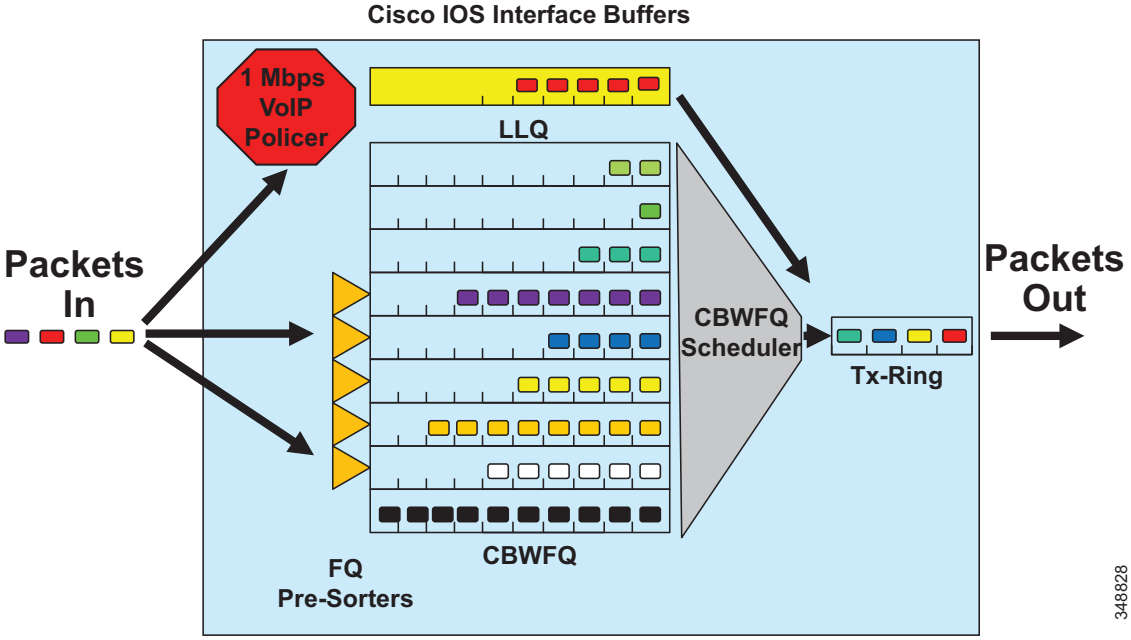
The depth of a CBWFQ is defined by its queue limit, which varies according to link speeds and platforms. This queue limit can be modified with the **queue-limit** policy-map class configuration command. In some cases, such as provisioning (bursty) TelePresence traffic in a CBWFQ, it is recommended to increase the queue limit from the default value. This is discussed in more detail in the section on Weighted-Random Early Detect, page 3-44.

Older (pre-HQF and pre-12.4(20)T) versions of Cisco IOS software include a legacy feature that disallows LLQ/CBWFQ policies from being attached to an interface if those policies explicitly allocate more than 75% of the interface's bandwidth to non-default traffic classes. This was intended as a safety feature that would always allow the default class as well as control-traffic classes to receive adequate bandwidth, and it allowed provisioning for Layer 2 bandwidth overhead. This feature can be overridden by applying the **max-reserved-bandwidth** interface command, which takes as a parameter the total percentage of interface bandwidth that can be explicitly provisioned (typically this value is set to 100). However, if this safety feature is overridden, then it is highly recommended that the default class be explicitly assigned no less than 25% of the link's bandwidth.

Low-Latency Queuing

Low-Latency Queuing (LLQ) is essentially CBWFQ combined with a strict priority queue. Basic LLQ operation is illustrated in Figure 3-15.

Figure 3-15 Cisco IOS (Single) LLQ Operation



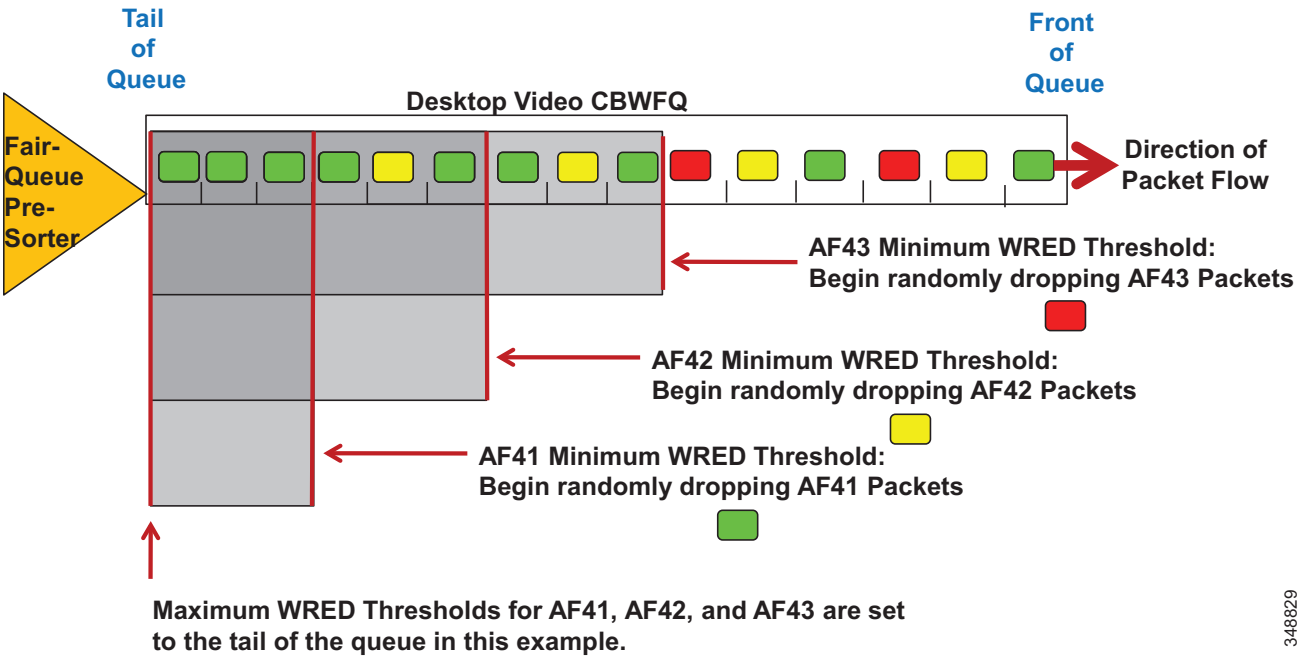
As shown in Figure 3-15, LLQ adds a strict-priority queue to the CBWFQ subsystem. The amount of bandwidth allocated to the LLQ is set by the **priority** policy-map class configuration command An interesting facet of Cisco IOS LLQ is the inclusion of an implicit policer that admits packets to the strict-priority queue. This implicit policer limits the bandwidth that can be consumed by servicing the real-time queue, and it thus prevents bandwidth starvation of the non-real-time flows serviced by the CBWFQ scheduler. The policing rate for this implicit policer is always set to match the bandwidth allocation of the strict-priority queue. If more traffic is offered to the LLQ class than it has been provisioned to accommodate, then the excess traffic will be dropped by the policer. And like the LLQ/CBWFQ systems, the implicit policer is active only during the event of congestion (as signaled to the Cisco IOS software by means of a full Tx-Ring).

Weighted-Random Early Detect

While congestion management mechanisms such as LLQ/CBWFQ manage the front of the queue, congestion avoidance mechanisms such as Weighted-Random Early Detect (WRED) manage the tail of the queue. Congestion avoidance mechanisms work best with TCP-based applications because selective dropping of packets causes the TCP windowing mechanisms to "throttle-back" and adjust the rate of flows to manageable rates.

The primary congestion avoidance mechanism in Cisco IOS is WRED, which randomly drops packets as queues fill to capacity. However, the randomness of this selection can be skewed by traffic weights. The weight can be IP Precedence (IPP) values, as is the case with default WRED which drops lower IPP values more aggressively (for example, statistically IPP 1 would be dropped more aggressively than IPP 6), or the weights can be AF Drop Precedence values, as is the case with DSCP-based WRED which statistically drops higher AF Drop Precedence values more aggressively (for example, AF43 is dropped more aggressively than AF42, which in turn is dropped more aggressively than AF41). DSCP-based WRED is enabled with the **dscp** keyword in conjunction with the **random-detect** policy-map class configuration command. The operation of DSCP-based WRED is illustrated in Figure 3-16.

Figure 3-16 Cisco IOS DSCP-Based WRED Operation



As shown in Figure 3-16, packets marked with a given Drop Precedence (AF43, AF42, or AF41) will begin to be dropped only when the queue fills beyond the minimum WRED threshold for the Drop Precedence value. Packets are always dropped randomly, but their probability of being dropped increases as the queue fills nearer the maximum WRED threshold for the Drop Precedence value. The maximum WRED thresholds are typically set at 100% (the tail of the queue), as shown in Figure 3-16; but the thresholds are configurable, and some advanced administrators may tune these WRED thresholds according to their needs, constraints, and preferences.

Additionally, the WRED thresholds on the AF class may be optimized. By default the minimum WRED thresholds for each AF class are 24, 28, and 32 packets for Drop-Precedence values 3, 2, and 1 respectively. These thresholds represent 60%, 70%, and 80% respectively of the default queue-depth of 64 packets. Also, by default the maximum WRED thresholds are set to 40 packets for all Drop-Precedence values for each AF class. Considering that the default queue-limit or depth is 64 packets, these default settings are inefficient on links experiencing sustained congestion that can cause a queue-depth of 40 packets (at which point all code points will be tail-dropped, despite the queue having the capacity to accommodate another 24 packets). Thus, an administrator may choose to tune these WRED thresholds so that each AF class has a minimum WRED threshold of 40, 45, and 50 packets for Drop-Precedence values 3, 2, and 1 respectively, which represent approximately 60%, 70%, and 80% of the default queue-depth of 64 packets, and/or the administrator may choose to tune the maximum WRED thresholds for each Drop-Precedence value for each AF class to the default queue-depth of 64 packets.

An example design is presented in the chapter on Bandwidth Management, page 13-1.

Considerations for Lower-Speed Links

Before placing voice and video traffic on a network, it is important to ensure that there is adequate bandwidth for all required applications. Once this bandwidth has been provisioned, voice priority queuing must be performed on all interfaces. This queuing is required to reduce jitter and possible packet loss if a burst of traffic oversubscribes a buffer. This queuing requirement is similar to the one for the LAN infrastructure.

Next, the WAN typically requires additional mechanisms such as traffic shaping to ensure that WAN links are not sent more traffic than they can handle, which could cause dropped packets.

Finally, link efficiency techniques can be applied to WAN paths. For example, link fragmentation and interleaving (LFI) can be used to prevent small voice packets from being queued behind large data packets, which could lead to unacceptable delays on low-speed links.

The goal of these QoS mechanisms is to ensure reliable, high-quality voice by reducing delay, packet loss, and jitter for the voice traffic. Table 3-5 lists the QoS features and tools required for the WAN infrastructure to achieve this goal based on the WAN link speed.

Table 3-5 QoS Features and Tools Required to Support Unified Communications for Each WAN Technology and Link Speed

WAN Technology	Link Speed: 56 kbps to 768 kbps	Link Speed: Greater than 768 kbps
Leased Lines	• Multilink Point-to-Point Protocol (MLP) • MLP Link Fragmentation and Interleaving (LFI) • Low Latency Queuing (LLQ) • Optional: Compressed Real-Time Transport Protocol (cRTP)	• LLQ
Frame Relay (FR)	• Traffic Shaping • LFI (FRF.12) • LLQ • Optional: cRTP • Optional: Voice-Adaptive Traffic Shaping (VATS) • Optional: Voice-Adaptive Fragmentation (VAF)	• Traffic Shaping • LLQ • Optional: VATS
Asynchronous Transfer Mode (ATM)	• TX-ring buffer changes • MLP over ATM • MLP LFI • LLQ • Optional: cRTP (requires MLP)	• TX-ring buffer changes • LLQ
Frame Relay and ATM Service Inter-Working (SIW)	• TX-ring buffer changes • MLP over ATM and FR • MLP LFI • LLQ • Optional: cRTP (requires MLP)	• TX-ring buffer changes • MLP over ATM and FR • LLQ
Multiprotocol Label Switching (MPLS)	• Same as above, according to the interface technology • Class-based marking is generally required to re-mark flows according to service provider specifications	• Same as above, according to the interface technology • Class-based marking is generally required to re-mark flows according to service provider specifications

The following sections highlight some of the most important features and techniques to consider when designing a WAN to support voice, video, and data traffic:

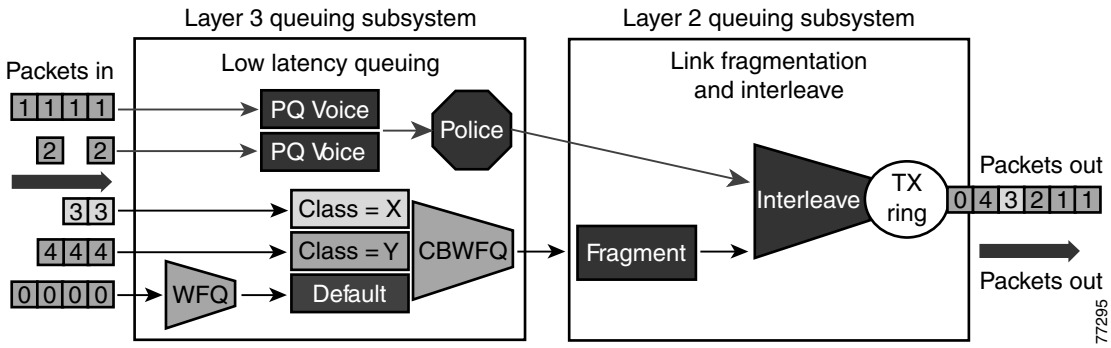
- Traffic Prioritization, page 3-47
- Link Efficiency Techniques, page 3-48
- Traffic Shaping, page 3-50

Traffic Prioritization

In choosing from among the many available prioritization schemes, the major factors to consider include the type of traffic involved and the type of media on the WAN. For multi-service traffic over an IP WAN, Cisco recommends low-latency queuing (LLQ) for all links. This method supports up to 64 traffic classes, with the ability to specify, for example, priority queuing behavior for voice and interactive video, minimum bandwidth class-based weighted fair queuing for voice control traffic, additional minimum bandwidth weighted fair queues for mission critical data, and a default best-effort queue for all other traffic types.

Figure 3-17 shows an example prioritization scheme.

Figure 3-17 Optimized Queuing for VoIP over the WAN



Cisco recommends the following prioritization criteria for LLQ:

- The criterion for *voice* to be placed into a priority queue is a DSCP value of 46 (EF).
- The criterion for *video conferencing* traffic to be placed into a class-based weighted fair queue (CBWFQ) is a DSCP value of 34 (AF41). Due to the larger packet sizes of video traffic, link speeds below 768 Kbps require packet fragmentation, which can happen only when video is placed in a separate CBWFQ. Video in a priority queue (PQ) is not fragmented.
- As the WAN links become congested, it is possible to starve the *voice control* signaling protocols, thereby eliminating the ability of the IP phones to complete calls across the IP WAN. Therefore, voice control protocols, such as H.323, MGCP, and Skinny Client Control Protocol (SCCP), require their own class-based weighted fair queue. The entrance criterion for this queue is a DSCP value of 24 (CS3).
- In some cases, certain data traffic might require better than best-effort treatment. This traffic is referred to as *mission-critical data*, and it is placed into one or more queues that have the required amount of bandwidth. The queuing scheme within this class is first-in-first-out (FIFO) with a minimum allocated bandwidth. Traffic in this class that exceeds the configured bandwidth limit is placed in the default queue. The entrance criterion for this queue could be a Transmission Control Protocol (TCP) port number, a Layer 3 address, or a DSCP/PHB value.
- All remaining enterprise traffic can be placed in a default queue for best-effort treatment. If you specify the keyword **fair**, the queuing algorithm will be weighted fair queuing (WFQ).

Scavenger Class

The Scavenger class is intended to provide less than best-effort services to certain applications. Applications assigned to this class have little or no contribution to the organizational objectives of the enterprise and are typically entertainment oriented in nature. Assigning Scavenger traffic to a minimal bandwidth queue forces it to be squelched to virtually nothing during periods of congestion, but it allows it to be available if bandwidth is not being used for business purposes, such as might occur during off-peak hours.

- Scavenger traffic should be marked as DSCP CS1.
- Scavenger traffic should be assigned the lowest configurable queuing service. For instance, in Cisco IOS, this means assigning a CBWFQ of 1% to Scavenger class.

Link Efficiency Techniques

The following link efficiency techniques improve the quality and efficiency of low-speed WAN links.

Compressed Real-Time Transport Protocol (cRTP)

You can increase link efficiency by using Compressed Real-Time Transport Protocol (cRTP). This protocol compresses a 40-byte IP, User Datagram Protocol (UDP), and RTP header into approximately two to four bytes. cRTP operates on a per-hop basis. Use cRTP on a particular link only if that link meets *all* of the following conditions:

- Voice traffic represents more than 33% of the load on the specific link.
- The link uses a low bit-rate codec (such as G.729).
- No other real-time application (such as video conferencing) is using the same link.

If the link fails to meet any one of the preceding conditions, then cRTP is not effective and you should not use it on that link. Another important parameter to consider before using cRTP is router CPU utilization, which is adversely affected by compression and decompression operations.

cRTP on ATM and Frame Relay Service Inter-Working (SIW) links requires the use of Multilink Point-to-Point Protocol (MLP).

Note that cRTP compression occurs as the final step before a packet leaves the egress interface; that is, after LLQ class-based queueing has occurred. Beginning in Cisco IOS Release 12.(2)2T and later, cRTP provides a feedback mechanism to the LLQ class-based queueing mechanism that allows the bandwidth in the *voice* class to be configured based on the compressed packet value. With Cisco IOS releases prior to 12.(2)2T, this mechanism is not in place, so the LLQ is unaware of the compressed bandwidth and, therefore, the *voice* class bandwidth has to be provisioned as if no compression is taking place. Table 3-6 shows an example of the difference in *voice* class bandwidth configuration given a 512-kbps link with G.729 codec and a requirement for 10 calls.

Note that Table 3-6 assumes 24 kbps for non-cRTP G.729 calls and 10 kbps for cRTP G.729 calls. These bandwidth numbers are based on voice payload and IP/UDP/RTP headers only. They do not take into consideration Layer 2 header bandwidth. However, actual bandwidth provisioning should also include Layer 2 header bandwidth based on the type WAN link used.

Table 3-6 LLQ Voice Class Bandwidth Requirements for 10 Calls with 512 kbps Link Bandwidth and G.729 Codec

Cisco IOS Release	With cRTP Not Configured	With cRTP Configured
Prior to 12.2(2)T	240 kbps	240 kbps[1]
12.2(2)T or later	240 kbps	100 kbps

1. 140 kbps of unnecessary bandwidth must be configured in the LLQ *voice* class.

It should also be noted that, beginning in Cisco IOS Release 12.2(13)T, cRTP can be configured as part of the voice class with the Class-Based cRTP feature. This option allows cRTP to be specified within a class, attached to an interface via a service policy. This new feature provides compression statistics and bandwidth status via the **show policy interface** command, which can be very helpful in determining the offered rate on an interface service policy class given the fact that cRTP is compressing the IP/RTP headers.

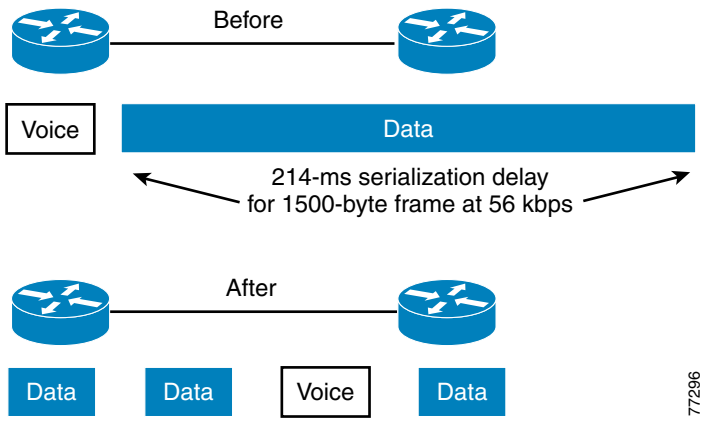
For additional recommendations about using cRTP with a Voice and Video Enabled IPSec VPN (V3PN), refer to the V3PN documentation available at

https://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns817/landing_voice_video.html

Link Fragmentation and Interleaving (LFI)

For low-speed links (less than 768 kbps), use of link fragmentation and interleaving (LFI) mechanisms is required for acceptable voice quality. This technique limits jitter by preventing voice traffic from being delayed behind large data frames, as illustrated in Figure 3-18. The two techniques that exist for this purpose are Multilink Point-to-Point Protocol (MLP) LFI (for Leased Lines, ATM, and SIW) and FRF.12 for Frame Relay.

Figure 3-18 Link Fragmentation and Interleaving (LFI)



Voice-Adaptive Fragmentation (VAF)

In addition to the LFI mechanisms mentioned above, voice-adaptive fragmentation (VAF) is another LFI mechanism for Frame Relay links. VAF uses FRF.12 Frame Relay LFI; however, once configured, fragmentation occurs only when traffic is present in the LLQ priority queue or when H.323 signaling packets are detected on the interface. This method ensures that, when voice traffic is being sent on the

WAN interface, large packets are fragmented and interleaved. However, when voice traffic is not present on the WAN link, traffic is forwarded across the link unfragmented, thus reducing the overhead required for fragmentation.

VAF is typically used in combination with voice-adaptive traffic shaping (see Voice-Adaptive Traffic Shaping (VATS), page 3-51). VAF is an optional LFI tool, and you should exercise care when enabling it because there is a slight delay between the time when voice activity is detected and the time when the LFI mechanism engages. In addition, a configurable deactivation timer (default of 30 seconds) must expire after the last voice packet is detected and before VAF is deactivated, so during that time LFI will occur unnecessarily. VAF is available in Cisco IOS Release 12.2(15)T and later.

Traffic Shaping

Traffic shaping is required for multiple-access, non-broadcast media such as ATM and Frame Relay, where the physical access speed varies between two endpoints and several branch sites are typically aggregated to a single router interface at the central site.

Figure 3-19 illustrates the main reasons why traffic shaping is needed when transporting voice and data on the same IP WAN.

Figure 3-19 Traffic Shaping with Frame Relay and ATM

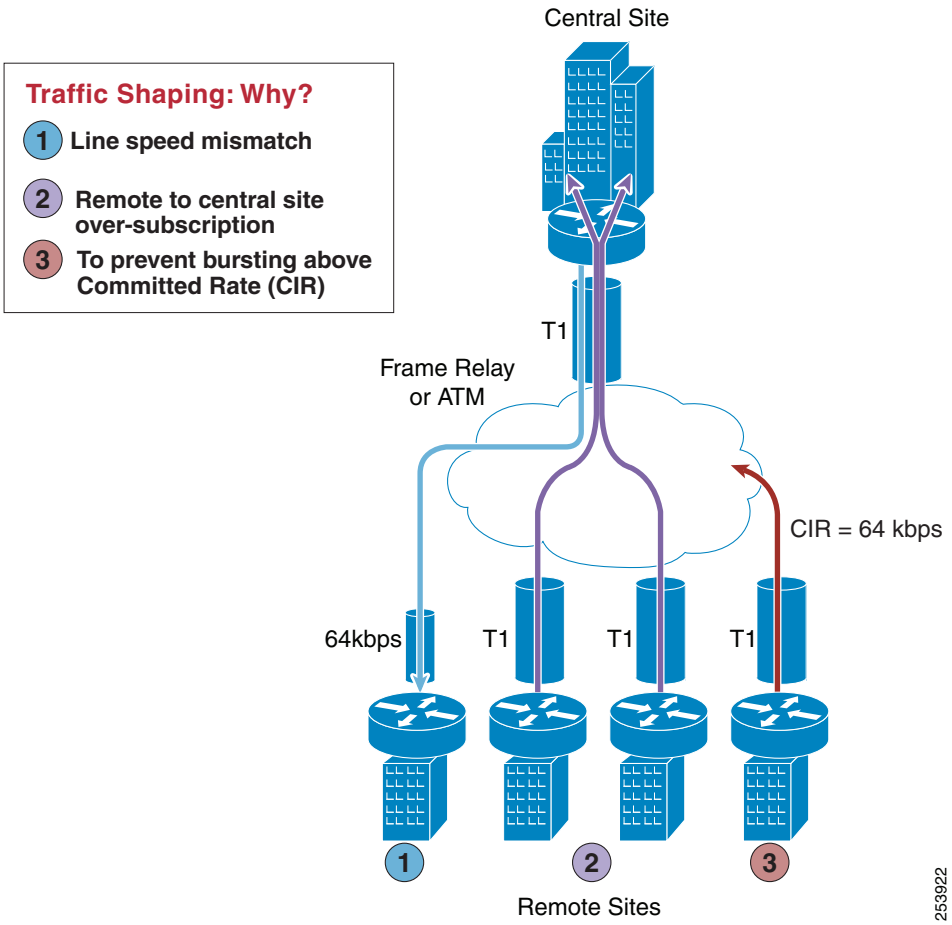


Figure 3-19 shows three different scenarios:

1. Line speed mismatch

   While the central-site interface is typically a high-speed one (such as T1 or higher), smaller remote branch interfaces may have significantly lower line speeds, such as 64 kbps. If data is sent at full rate from the central site to a slow-speed remote site, the interface at the remote site might become congested, resulting in dropped packets which causes a degradation in voice quality.

2. Oversubscription of the link between the central site and the remote sites

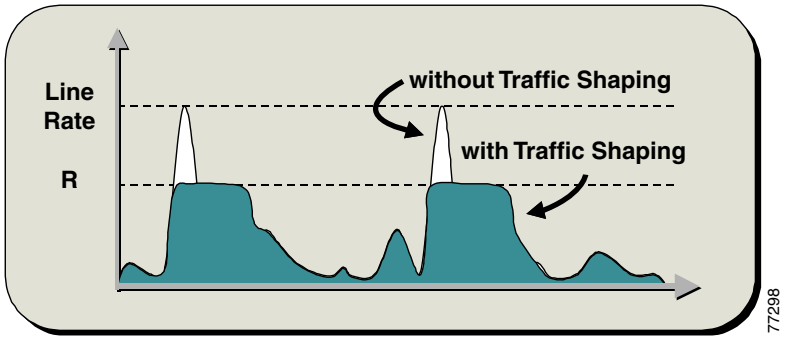
   It is common practice in Frame Relay or ATM networks to oversubscribe bandwidth when aggregating many remote sites to a single central site. For example, there may be multiple remote sites that connect to the WAN with a T1 interface, yet the central site has only a single T1 interface. While this configuration allows the deployment to benefit from statistical multiplexing, the router interface at the central site can become congested during traffic bursts, thus degrading voice quality.

3. Bursting above Committed Information Rate (CIR)

   Another common configuration is to allow traffic bursts above the CIR, which represents the rate that the service provider has guaranteed to transport across its network with no loss and low delay. For example, a remote site with a T1 interface might have a CIR of only 64 kbps. When more than 64 kbps worth of traffic is sent across the WAN, the provider marks the additional traffic as "discard eligible." If congestion occurs in the provider network, this traffic will be dropped with no regard to traffic classification, possibly having a negative effect on voice quality.

Traffic shaping provides a solution to these issues by limiting the traffic sent out an interface to a rate lower than the line rate, thus ensuring that no congestion occurs on either end of the WAN. Figure 3-20 illustrates this mechanism with a generic example, where R is the rate with traffic shaping applied.

Figure 3-20 Traffic Shaping Mechanism



Voice-Adaptive Traffic Shaping (VATS)

VATS is an optional dynamic mechanism that shapes traffic on Frame Relay permanent virtual circuits (PVCs) at different rates based on whether voice is being sent across the WAN. The presence of traffic in the LLQ voice priority queue or the detection of H.323 signaling on the link causes VATS to engage. Typically, Frame Relay shapes traffic to the guaranteed bandwidth or CIR of the PVC at all times. However, because these PVCs are typically allowed to burst above the CIR (up to line speed), traffic shaping keeps traffic from using the additional bandwidth that might be present in the WAN. With VATS enabled on Frame Relay PVCs, WAN interfaces are able to send at CIR when voice traffic is present on the link. However, when voice is not present, non-voice traffic is able to burst up to line speed and take advantage of the additional bandwidth that might be present in the WAN.

When VATS is used in combination with voice-adaptive fragmentation (VAF) (see Link Fragmentation and Interleaving (LFI), page 3-49), all non-voice traffic is fragmented and all traffic is shaped to the CIR of the WAN link when voice activity is detected on the interface.

As with VAF, exercise care when enabling VATS because activation can have an adverse effect on non-voice traffic. When voice is present on the link, data applications will experience decreased throughput because they are throttled back to well below CIR. This behavior will likely result in packet drops and delays for non-voice traffic. Furthermore, after voice traffic is no longer detected, the deactivation timer (default of 30 seconds) must expire before traffic can burst back to line speed. It is important, when using VATS, to set end-user expectations and make them aware that data applications will experience slowdowns on a regular basis due to the presence of voice calls across the WAN. VATS is available in Cisco IOS Release 12.2(15)T and later.

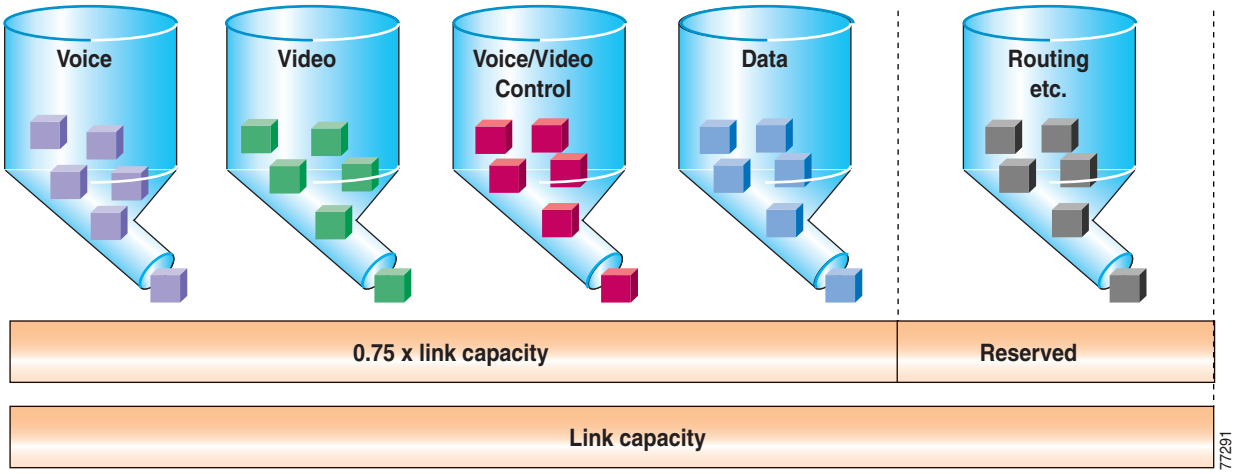
For more information on the Voice-Adaptive Traffic Shaping and Fragmentation features and how to configure them, refer to the documentation at

https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/wan_frly/configuration/15-mt/wan-frly-15-mt-book.html

Bandwidth Provisioning

Properly provisioning the network bandwidth is a major component of designing a successful IP network. You can calculate the required bandwidth by adding the bandwidth requirements for each major application (for example, voice, video, and data). This sum then represents the minimum bandwidth requirement for any given link, and it should not exceed approximately 75% of the total available bandwidth for the link. This 75% rule assumes that some bandwidth is required for overhead traffic, such as routing and Layer 2 keep-alives. Figure 3-21 illustrates this bandwidth provisioning process.

Figure 3-21 Link Bandwidth Provisioning



In addition to using no more than 75% of the total available bandwidth for data, voice, and video, the total bandwidth configured for all LLQ priority queues should typically not exceed 33% of the total link bandwidth. Provisioning more than 33% of the available bandwidth for the priority queue can be problematic for a number of reasons. First, provisioning more than 33% of the bandwidth for voice can result in increased CPU usage. Because each voice call will send 50 packets per second (with 20 ms samples), provisioning for large numbers of calls in the priority queue can lead to high CPU levels due to high packet rates. In addition, if more than one type of traffic is provisioned in the priority queue (for

example, voice and video), this configuration defeats the purpose of enabling QoS because the priority queue essentially becomes a first-in, first-out (FIFO) queue. A larger percentage of reserved priority bandwidth effectively dampens the QoS effects by making more of the link bandwidth FIFO. Finally, allocating more than 33% of the available bandwidth can effectively starve any data queues that are provisioned. Obviously, for very slow links (less than 192 kbps), the recommendation to provision no more than 33% of the link bandwidth for the priority queue(s) might be unrealistic because a single call could require more than 33% of the link bandwidth. In these situations, and in situations where specific business needs cannot be met while holding to this recommendation, it may be necessary to exceed the 33% rule.

From a traffic standpoint, an IP telephony call consists of two parts:

- The voice and video bearer streams, which consists of Real-Time Transport Protocol (RTP) packets that contain the actual voice samples.
- The call control signaling, which consists of packets belonging to one of several protocols, according to the endpoints involved in the call (for example, H.323, MGCP, SCCP, or (J)TAPI). Call control functions are, for instance, those used to set up, maintain, tear down, or redirect a call.

Bandwidth provisioning should include not only the bearer traffic but also the call control traffic. In fact, in multisite WAN deployments, the call control traffic (as well as the bearer traffic) must traverse the WAN, and failure to allocate sufficient bandwidth for it can adversely affect the user experience.

The next three sub-sections describe the bandwidth provisioning recommendations for the following types of traffic:

- Voice and video bearer traffic in all multisite WAN deployments (see Provisioning for Bearer Traffic, page 3-53)
- Call control traffic in multisite WAN deployments with centralized call processing (see Provisioning for Call Control Traffic with Centralized Call Processing, page 3-57)
- Call control traffic in multisite WAN deployments with distributed call processing (see Provisioning for Call Control Traffic with Distributed Call Processing, page 3-61)

Provisioning for Bearer Traffic

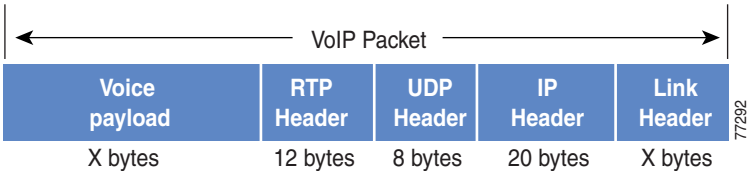
The section describes bandwidth provisioning for the following types of traffic:

- Voice Bearer Traffic, page 3-53
- Video Bearer Traffic, page 3-56

Voice Bearer Traffic

As illustrated in Figure 3-22, a voice-over-IP (VoIP) packet consists of the voice payload, IP header, User Datagram Protocol (UDP) header, Real-Time Transport Protocol (RTP) header, and Layer 2 Link header. When Secure Real-Time Transport Protocol (SRTP) encryption is used, the voice payload for each packet is increased by 4 bytes. The link header varies in size according to the Layer 2 media used.

Figure 3-22 Typical VoIP Packet



The bandwidth consumed by VoIP streams is calculated by adding the packet payload and all headers (in bits), then multiplying by the packet rate per second, as follows:

Layer 2 bandwidth in kbps = [(Packets per second) * (*X* bytes for voice payload + 40 bytes for RTP/UDP/IP headers + *Y* bytes for Layer 2 overhead) * 8 bits] / 1000

Layer 3 bandwidth in kbps = [(Packets per second) * (*X* bytes for voice payload + 40 bytes for RTP/UDP/IP headers) * 8 bits] / 1000

Packets per second = [1/(sampling rate in msec)] * 1000

Voice payload in bytes = [(codec bit rate in kbps) * (sampling rate in msec)] / 8

Table 3-7 details the Layer 3 bandwidth per VoIP flow. Table 3-7 lists the bandwidth consumed by the voice payload and IP header only, at a default packet rate of 50 packets per second (pps) and at a rate of 33.3 pps for both non-encrypted and encrypted payloads. Table 3-7 does not include Layer 2 header overhead and does not take into account any possible compression schemes, such as compressed Real-Time Transport Protocol (cRTP). You can use the Service Parameters menu in Unified CM Administration to adjust the codec sampling rate.

Table 3-7 Bandwidth Consumption for Voice Payload and IP Header Only

CODEC	Sampling Rate	Voice Payload in Bytes	Packets per Second	Bandwidth per Conversation
G.711 and G.722-64k	20 ms	160	50.0	80.0 kbps
G.711 and G.722-64k (SRTP)	20 ms	164	50.0	81.6 kbps
G.711 and G.722-64k	30 ms	240	33.3	74.7 kbps
G.711 and G.722-64k (SRTP)	30 ms	244	33.3	75.8 kbps
iLBC	20 ms	38	50.0	31.2 kbps
iLBC (SRTP)	20 ms	42	50.0	32.8 kbps
iLBC	30 ms	50	33.3	24.0 kbps
iLBC (SRTP)	30 ms	54	33.3	25.1 kbps
G.729A	20 ms	20	50.0	24.0 kbps
G.729A (SRTP)	20 ms	24	50.0	25.6 kbps
G.729A	30 ms	30	33.3	18.7 kbps
G.729A (SRTP)	30 ms	34	33.3	19.8 kbps

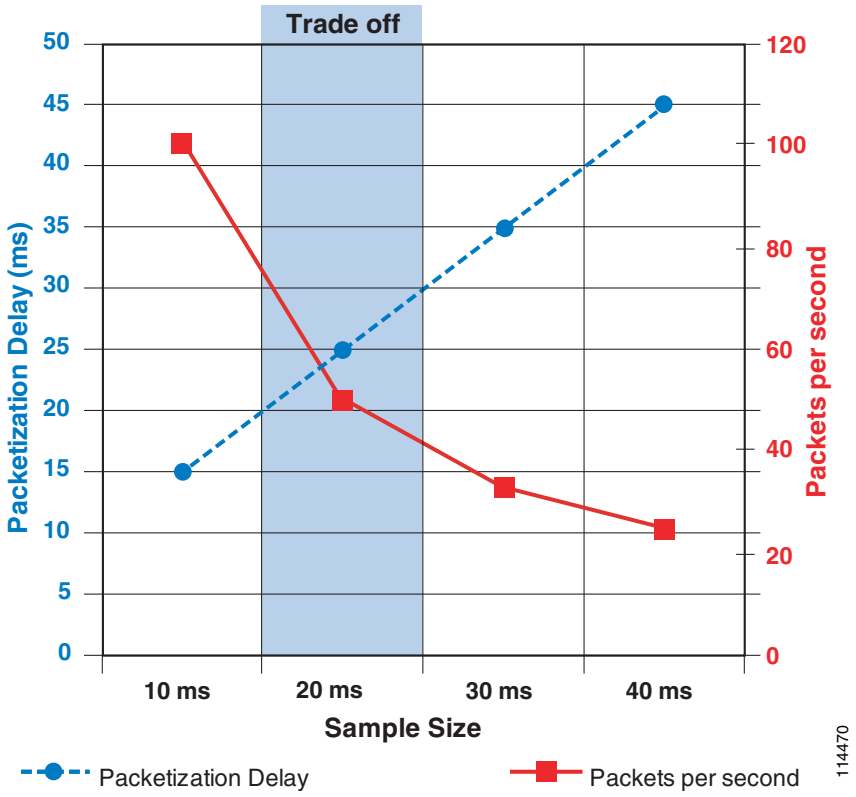
A more accurate method for provisioning is to include the Layer 2 headers in the bandwidth calculations. Table 3-8 lists the amount of bandwidth consumed by voice traffic when the Layer 2 headers are included in the calculations.

Table 3-8 Bandwidth Consumption with Layer 2 Headers Included

CODEC	Header Type and Size						
	Ethernet 14 Bytes	PPP 6 Bytes	ATM 53-Byte Cells with a 48-Byte Payload	Frame Relay 4 Bytes	MLPPP 10 Bytes	MPLS 4 Bytes	WLAN 24 Bytes
G.711 and G.722-64k at 50.0 pps	85.6 kbps	82.4 kbps	106.0 kbps	81.6 kbps	84.0 kbps	81.6 kbps	89.6 kbps
G.711 and G.722-64k (SRTP) at 50.0 pps	87.2 kbps	84.0 kbps	106.0 kbps	83.2 kbps	85.6 kbps	83.2 kbps	N/A
G.711 and G.722-64k at 33.3 pps	78.4 kbps	76.3 kbps	84.8 kbps	75.7 kbps	77.3 kbps	75.7 kbps	81.1 kbps
G.711 and G.722-64k (SRTP) at 33.3 pps	79.5 kbps	77.4 kbps	84.8 kbps	76.8 kbps	78.4 kbps	76.8 kbps	N/A
iLBC at 50.0 pps	36.8 kbps	33.6 kbps	42.4 kbps	32.8 kbps	35.2 kbps	32.8 kbps	40.8 kbps
iLBC (SRTP) at 50.0 pps	38.4 kbps	35.2 kbps	42.4 kbps	34.4 kbps	36.8 kbps	34.4 kbps	42.4 kbps
iLBC at 33.3 pps	27.7 kbps	25.6 kbps	28.3 kbps	25.0 kbps	26.6 kbps	25.0 kbps	30.4 kbps
iLBC (SRTP) at 33.3 pps	28.8 kbps	26.6 kbps	42.4 kbps	26.1 kbps	27.7 kbps	26.1 kbps	31.5 kbps
G.729A at 50.0 pps	29.6 kbps	26.4 kbps	42.4 kbps	25.6 kbps	28.0 kbps	25.6 kbps	33.6 kbps
G.729A (SRTP) at 50.0 pps	31.2 kbps	28.0 kbps	42.4 kbps	27.2 kbps	29.6 kbps	27.2 kbps	35.2 kbps
G.729A at 33.3 pps	22.4 kbps	20.3 kbps	28.3 kbps	19.7 kbps	21.3 kbps	19.8 kbps	25.1 kbps
G729A (SRTP) at 33.3 pps	23.5 kbps	21.4 kbps	28.3 kbps	20.8 kbps	22.4 kbps	20.8 kbps	26.2 kbps

While it is possible to configure the sampling rate above 30 ms, doing so usually results in very poor voice quality. As illustrated in Figure 3-23, as sampling size increases, the number of packets per second decreases, resulting in a smaller impact to the CPU of the device. Likewise, as the sample size increases, IP header overhead is lower because the payload per packet is larger. However, as sample size increases, so does packetization delay, resulting in higher end-to-end delay for voice traffic. The trade-off between packetization delay and packets per second must be considered when configuring sample size. While this trade-off is optimized at 20 ms, 30 ms sample sizes still provide a reasonable ratio of delay to packets per second; however, with 40 ms sample sizes, the packetization delay becomes too high.

Figure 3-23 Voice Sample Size: Packets per Second vs. Packetization Delay



Video Bearer Traffic

For audio, it is relatively easy to calculate a percentage of overhead per packet given the sample size of each packet. For video, however, it is nearly impossible to calculate an exact percentage of overhead because the payload varies depending upon how much motion is present in the video (that is, how many pixels changed since the last frame).

To resolve this inability to calculate the exact overhead ratio for video, Cisco recommends that you add 20% to the call speed regardless of which type of Layer-2 medium the packets are traversing. The additional 20% gives plenty of headroom to allow for the differences between Ethernet, ATM, Frame Relay, PPP, HDLC, and other transport protocols, as well as some cushion for the bursty nature of video traffic.

Note that the call speed requested by the endpoint (for example, 128 kbps, 256 kbps, and so forth) represents the maximum burst speed of the call, with some additional amount for a cushion. The average speed of the call is typically much less than these values.

Provisioning for Call Control Traffic

When Unified Communications endpoints are separated from their call control application by a WAN, or when two interconnected Unified Communications systems are separated by a WAN, consideration must be given to the amount of bandwidth that must be provisioned for call control and signaling traffic between these endpoints and systems. This section discusses WAN bandwidth provisioning for call signaling traffic where centralized or distributed call processing models are deployed. For more information on Unified Communications centralized and distributed call processing deployment models, see Collaboration Deployment Models, page 10-1.

Provisioning for Call Control Traffic with Centralized Call Processing

In a centralized call processing deployment, the Unified CM cluster and the applications (such as voicemail) are located at the central site, while several remote sites are connected through an IP WAN. The remote sites rely on the centralized Unified CMs to handle their call processing.

The following considerations apply to this deployment model:

- Each time a remote branch phone places a call, the control traffic traverses the IP WAN to reach the Unified CM at the central site, even if the call is local to the branch.
- The signaling protocols that may traverse the IP WAN in this deployment model are SCCP (encrypted and non-encrypted), SIP (encrypted and non-encrypted), H.323, MGCP, and CTI-QBE. All the control traffic is exchanged between a Unified CM at the central site and endpoints or gateways at the remote branches.

As a consequence, you must provision bandwidth for control traffic that traverses the WAN between the branch routers and the WAN aggregation router at the central site.

The control traffic that traverses the WAN in this scenario can be split into two categories:

- Quiescent traffic, which consists of keep-alive messages periodically exchanged between the branch endpoints (phones and gateways) and Unified CM, regardless of call activity. This traffic is a function of the quantity of endpoints.
- Call-related traffic, which consists of signaling messages exchanged between the branch endpoints and the Unified CM at the central site when a call needs to be set up, torn down, forwarded, and so forth. This traffic is a function of the quantity of endpoints and their associated call volume.

To obtain an estimate of the generated call control traffic, it is necessary to make some assumptions regarding the average number of calls per hour made by each branch IP phone. In the interest of simplicity, the calculations in this section assume an average of 10 calls per hour per phone.

Note If this average number does not satisfy the needs of your specific deployment, you can calculate the recommended bandwidth by using the advanced formulas provided in Advanced Formulas, page 3-59.

Given the assumptions made, and initially considering the case of a remote branch with no signaling encryption configured, the recommended bandwidth needed for call control traffic can be obtained from the following formula:

Equation 1A: Recommended Bandwidth Needed for SCCP Control Traffic without Signaling Encryption.

Bandwidth (bps) = 265 * (Number of IP phones and gateways in the branch)

Equation 1B: Recommended Bandwidth Needed for SIP Control Traffic without Signaling Encryption.

Bandwidth (bps) = 538 * (Number of IP phones and gateways in the branch)

If a site features a mix of SCCP and SIP endpoints, the two equations above should be employed separately for the quantity of each type of phone used, and the results added.

Equation 1 and all other formulas within this section include a 25% over-provisioning factor. Control traffic has a bursty nature, with peaks of high activity followed by periods of low activity. For this reason, assigning just the minimum bandwidth required to a control traffic queue can result in undesired effects such as buffering delays and, potentially, packet drops during periods of high activity. The default queue depth for a Class-Based Weighted Fair Queuing (CBWFQ) queue in Cisco IOS equals 64 packets. The bandwidth assigned to this queue determines its servicing rate. Assuming that the bandwidth configured is the average bandwidth consumed by this type of traffic, it is clear that, during the periods of high activity, the servicing rate will not be sufficient to "drain" all the incoming packets out of the queue, thus causing them to be buffered. Note that, if the 64-packet limit is reached, any subsequent packets are either assigned to the best-effort queue or are dropped. It is therefore advisable to introduce this 25% over-provisioning factor to absorb and smooth the variations in the traffic pattern and to minimize the risk of a temporary buffer overrun. This is equivalent to increasing the servicing rate of the queue.

If encryption is configured, the recommended bandwidth is affected because encryption increases the size of signaling packets exchanged between Unified CM and the endpoints. The following formula takes into account the impact of signaling encryption:

Equation 2A: Recommended Bandwidth Needed for SCCP Control Traffic with Signaling Encryption.

Bandwidth with signaling encryption (bps) = 415 * (Number of IP phones and gateways in the branch)

Equation 2B: Recommended Bandwidth Needed for SIP Control Traffic with Signaling Encryption.

Bandwidth with signaling encryption (bps) = 619 * (Number of IP phones and gateways in the branch)

If we now take into account the fact that the smallest bandwidth that can be assigned to a queue on a Cisco IOS router is 8 kbps, we can summarize the values of minimum and recommended bandwidth for various branch office sizes, as shown in Table 3-9.

Table 3-9 Recommended Layer 3 Bandwidth for Call Control Traffic With and Without Signaling Encryption

Branch Office Size (Number of IP Phones and Gateways)	Recommended Bandwidth for SCCP Control Traffic (no encryption)	Recommended Bandwidth for SCCP Control Traffic (with encryption)	Recommended Bandwidth for SIP Control Traffic (no encryption)	Recommended Bandwidth for SIP Control Traffic (with encryption)
1 to 10	8 kbps	8 kbps	8 kbps	8 kbps
20	8 kbps	9 kbps	11 kbps	12 kbps
30	8 kbps	13 kbps	16 kbps	19 kbps
40	11 kbps	17 kbps	22 kbps	25 kbps
50	14 kbps	21 kbps	27 kbps	31 kbps
100	27 kbps	42 kbps	54 kbps	62 kbps
150	40 kbps	62 kbps	81 kbps	93 kbps

The values in Table 3-9 are out-of-date for newer models of phones running SIP signaling with more features and functions that can require additional signaling overhead within the SIP stack. Also, the calculations in the above equations assume basic single line calls. Therefore, we recommend monitoring and testing usage of the signaling queue to determine what adjustments are needed if there are queue tail drops during congested times of the day. On higher bandwidth WAN links, configuring a greater value

for the queue is recommended, since any unused bandwidth will become available for all other queues on the WAN. Therefore, it is best to use the values for SIP with and without encryption in Table 3-9 as a guide, and adjust to higher values per phone.

Note The above recommendation is for WAN queuing bandwidth configuration and not for LAN access port policing configuration. For LAN access port policing we recommend setting the value to 80 kbps or more, depending on the expected signaling spikes from various use cases. (Older documents recommend 32 kbps, but this is no longer the norm for many signaling use cases). For example, a phone with a busy lamp field (BLF) configured on a line will generate a SIP NOTIFY and a SIP 200OK for each busy indication. Thus, a phone with a large number of associated BLFs could cause a spike in SIP signaling, as could going on and off hook quickly multiple times in one second. Therefore, we recommend accounting for your worst case scenario for signaling before policing on an access port switch.

Advanced Formulas

The previous formulas presented in this section assume an average call rate per phone of 10 calls per hour. However, this rate might not correspond to your deployment if the call patterns are significantly different (for example, with call center agents at the branches). To calculate call control bandwidth requirements in these cases, use the following formulas, which contain an additional variable (CH) that represents the average calls per hour per phone:

Equation 3A: Recommended Bandwidth Needed for SCCP Control Traffic for a Branch with No Signaling Encryption.

Bandwidth (bps) = (53 + 21 * CH) * (Number of IP phones and gateways in the branch)

Equation 3B: Recommended Bandwidth Needed for SIP Control Traffic for a Branch with No Signaling Encryption.

Bandwidth (bps) = (138 + 40 * CH) * (Number of IP phones and gateways in the branch)

Equation 4A: Recommended Bandwidth Needed for SCCP Control Traffic for a Remote Branch with Signaling Encryption.

Bandwidth with signaling encryption (bps) = (73.5 + 33.9 * CH) * (Number of IP phones and gateways in the branch)

Equation 4B: Recommended Bandwidth Needed for SIP Control Traffic for a Remote Branch with Signaling Encryption.

Bandwidth with signaling encryption (bps) = (159 + 46 * CH) * (Number of IP phones and gateways in the branch)

Note Equations 3A and 4A are based on the default SCCP keep-alive period of 30 seconds, while equations 3B and 4B are based on the default SIP keep-alive period of 120 seconds.

Considerations for Shared Line Appearances

Calls placed to shared line appearances, or calls sent to line groups using the Broadcast distribution algorithm, have two net effects on the bandwidth consumed by the system:

- Because all the phones on which the line is configured ring simultaneously, they represent a load on the system corresponding to a much higher calls-per-hour (CH) value than the CH of the line. The corresponding bandwidth consumption is therefore increased. The network infrastructure's

bandwidth provisioning requires adjustments when WAN-connected shared line functionality is deployed. The CH value employed for Equations 3 and 4 must be increased according to the following formula:

CHS = CHL * (Number line appearances) / (Number of lines)

Where CHS is the shared-line calls per hour to be used in Equations 3 and 4, and CHL is the calls-per-hour rating of the line. For example, if a site is configured with 5 lines making an average of 6 calls per hour but 2 of those lines are shared across 4 different phones, then:

Number of lines = 5

Number of line appearances = (2 lines appear on 4 phones, and 3 lines appear on only one phone) = (2*4) + 3 = 11 line appearances

CHL = 6

CHS = 6 * (11 / 5) = 13.2

- Because each of the ringing phones requires a separate signaling control stream, the quantity of packets sent from Unified CM to the same branch is increased in linear proportion to the quantity of phones ringing. Because Unified CM is attached to the network through an interface that supports 100 Mbps or more, it can instantaneously generate a very large quantity of packets that must be buffered while the queuing mechanism is servicing the signaling traffic. The servicing speed is limited by the WAN interface's effective information transfer speed, which is typically two orders of magnitude smaller than 100 Mbps.

  This traffic may overwhelm the queue depth of the central site's WAN router. By default, the queue depth available for each of the classes of traffic in Cisco IOS is 64. In order to prevent any packets from being dropped before they are queued for the WAN interface, you must ensure that the signaling queue's depth is sized to hold all the packets from at least one full shared-line event for each shared-line phone. Avoiding drops is paramount in ensuring that the call does not create a race condition where dropped packets are retransmitted, causing system response times to suffer.

  Therefore, the quantity of packets required to operate shared-line phones is as follows:

  - SCCP protocol: 13 packets per shared-line phone
  - SIP protocol: 11 packets per shared-line phone

  For example, with SCCP and with 6 phones sharing the same line, the queue depth for the signaling class of traffic must be adjusted to a minimum of 78. Table 3-10 provides recommended queue depths based on the quantity of shared line appearances within a branch site.

Table 3-10 Recommended Queue Depth per Branch Site

Number of Shared Line Appearances	Queue Depth (Packets)	
	SCCP	SIP
5	65	55
10	130	110
15	195	165
20	260	220
25	325	275

When using a Layer 2 WAN technology such as Frame Relay, this adjustment must be made on the circuit corresponding to the branch where the shared-line phones are located.

When using a Layer 3 WAN technology such as MPLS, there may be a single signaling queue servicing multiple branches. In this case, adjustment must be made for the total of all branches serviced.

Provisioning for Call Control Traffic with Distributed Call Processing

In distributed call processing deployments, Unified CM Clusters, each following either the single-site model or the centralized call processing model, are connected through an IP WAN. The signaling protocol used to place a call across the WAN is SIP (H.323 trunks are no longer recommended between Unified CM clusters). This SIP protocol control traffic that traverses the WAN belongs to signaling traffic associated with a media stream, exchanged over an intercluster trunk when a call needs to be set up, torn down, forwarded, and so on.

Because the total amount of control traffic depends on the number of calls that are set up and torn down at any given time, it is necessary to make some assumptions about the call patterns and the link utilization. Using a traditional telephony analogy, we can view the portion of the WAN link that has been provisioned for voice and video as a number of *virtual tie lines* and derive the protocol signaling traffic associated with the virtual tie lines.

Assuming an average call duration of 2 minutes and 100 percent utilization of each virtual tie line, we can derive that each tie line carries a volume of 30 calls per hour. This assumption allows us to obtain the following formula that expresses the recommended bandwidth for call control traffic as a function of the number of virtual tie lines.

Equation 6: Recommended Bandwidth Based on Number of Virtual Tie Lines.

Recommended Bandwidth (bps) = 116 * (Number of virtual tie lines)

If we take into account the fact that 8 kbps is the smallest bandwidth that can be assigned to a queue on a Cisco IOS router, we can deduce that a minimum queue size of 8 kbps can accommodate the call control traffic generated by up to 70 virtual tie lines or 2,100 calls per hour. This amount of 8 kbps for SIP signaling traffic between clusters should be sufficient for most large enterprise deployments.

Wireless LAN Infrastructure

Wireless LAN infrastructure design becomes important when collaboration endpoints are added to the wireless LAN (WLAN) portions of a converged network. With the introduction of Cisco Unified Wireless endpoints, voice and video traffic has moved onto the WLAN and is now converged with the existing data traffic there. Just as with wired LAN and wired WAN infrastructure, the addition of voice and video in the WLAN requires following basic configuration and design best-practices for deploying a highly available network. In addition, proper WLAN infrastructure design requires understanding and deploying QoS on the wireless network to ensure end-to-end voice and video quality on the entire network. The following sections discuss these requirements:

- Architecture for Voice and Video over WLAN, page 3-62
- High Availability for Voice and Video over WLAN, page 3-66
- Capacity Planning for Voice and Video over WLAN, page 3-68
- Design Considerations for Voice and Video over WLAN, page 3-68

For more information about voice and video over wireless LANs, refer to the *Real-Time Traffic over Wireless LAN Solution Reference Network Design Guide*, available at

https://www.cisco.com/en/US/docs/solutions/Enterprise/Mobility/RToWLAN/CCVP_BK_R7805F20_00_rtowlan-srnd.html

Architecture for Voice and Video over WLAN

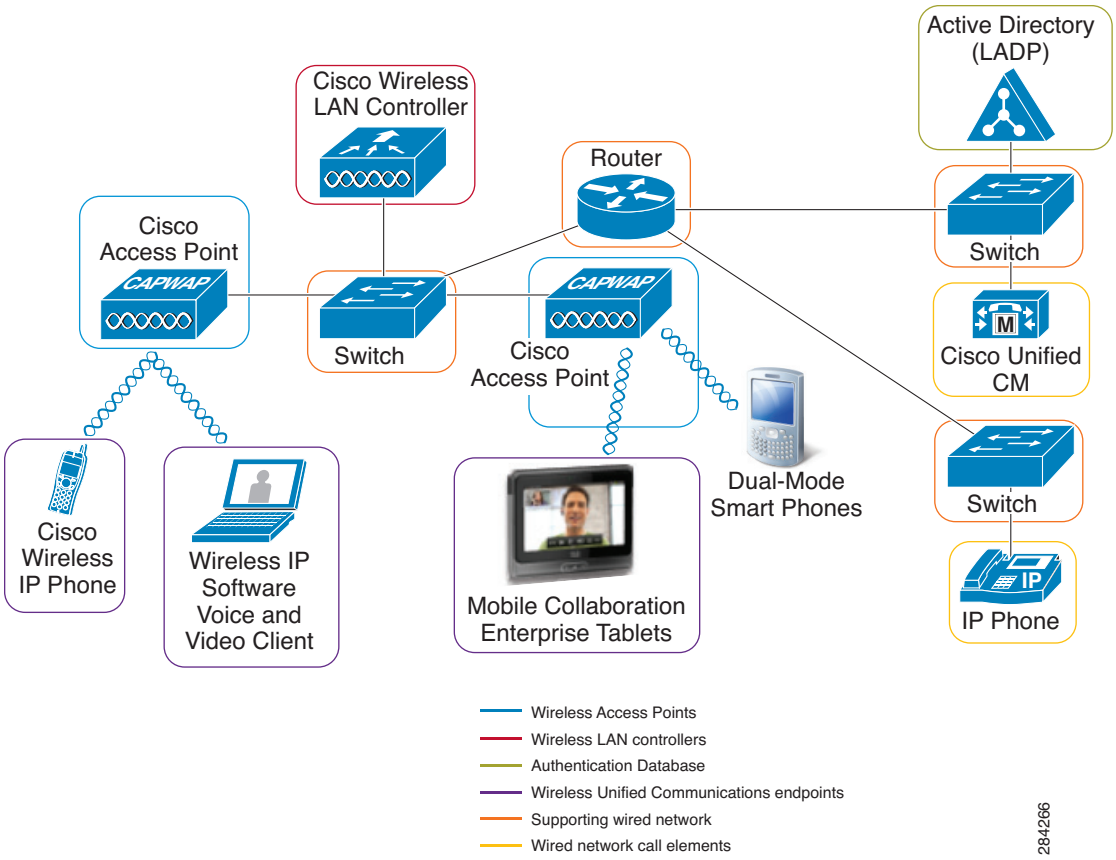
IP telephony architecture has used wired devices since its inception, but enterprise users have long sought the ability to communicate while moving through the company premises. Wireless IP networks have enabled IP telephony to deliver enterprise mobility by providing on-premises roaming communications to the users with wireless IP telephony devices.

Wireless IP telephony and wireless IP video telephony are extensions of their wired counterparts, which leverage the same call elements. Additionally, wireless IP telephony and IP video telephony take advantage of wireless 802.11-enabled media, thus providing a cordless IP voice and video experience. The cordless experience is achieved by leveraging the wireless network infrastructure elements for the transmission and reception of the control and media packets.

The architecture for voice and video over wireless LAN includes the following basic elements, illustrated in Figure 3-24:

- Wireless Access Points, page 3-63
- Wireless LAN Controllers, page 3-64
- Authentication Database, page 3-64
- Supporting Wired Network, page 3-64
- Wireless Collaboration Endpoints, page 3-65
- Wired Call Elements, page 3-65

Figure 3-24 Basic Layout for a Voice and Video Wireless Network

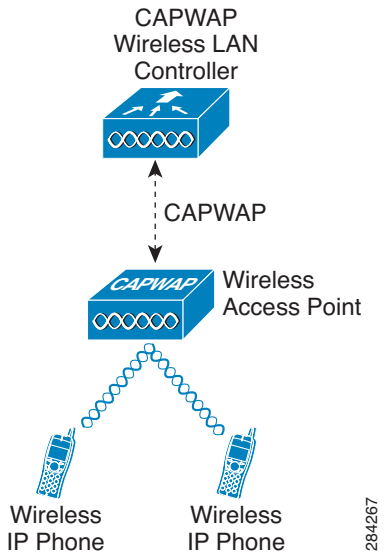


Wireless Access Points

The wireless access points enable wireless devices (Unified Communications endpoints in the case of voice and video over WLAN) to communicate with wired network elements. Access points function as adapters between the wired and wireless world, creating an entry-way between these two media. Cisco access points can be managed by a wireless LAN controller (WLC) or they can function in autonomous mode. When the access points are managed by a WLC they are referred as Lightweight Access Points, and in this mode they use the Lightweight Access Point Protocol (LWAPP) or Control and Provisioning of Wireless Access Points (CAPWAP) protocol, depending on the controller version, when communicating with the WLC.

Figure 3-25 illustrates the basic relationship between lightweight access points and WLCs. Although the example depicted in Figure 3-25 is for a CAPWAP WLC, from the traffic flow and relationship perspective there are no discernible differences between CAPWAP and LWAPP, so the example also applies to wireless LWAPP networks. Some advantages of leveraging WLCs and lightweight access points for the wireless infrastructure include ease of management, dynamic network tuning, and high availability. However, if you are using the managed mode instead of the autonomous mode in the access points, you need to consider the network tunneling effect of the LWAP-WLC communication architecture when designing your solution. This network tunneling effect is discussed in more depth in the section on Wireless LAN Controller Design Considerations, page 3-73.

Figure 3-25 Lightweight Access Point



Wireless LAN Controllers

Many corporate environments require deployment of wireless networks on a large scale. The wireless LAN controller (WLC) is a device that assumes a central role in the wireless network and helps to make it easier to manage such large-scale deployments. Traditional roles of access points, such as association or authentication of wireless clients, are done by the WLC. Access points, called Lightweight Access Points (LWAPs) in the Unified Communications environment, register themselves with a WLC and tunnel all the management and data packets to the WLCs, which then switch the packets between wireless clients and the wired portion of the network. All the configurations are done on the WLC. LWAPs download the entire configuration from WLCs and act as a wireless interface to the clients.

Authentication Database

The authentication database is a core component of the wireless networks, and it holds the credentials of the users to be authenticated while the wireless association is in progress. The authentication database provides the network administrators with a centralized repository to validate the credentials. Network administrators simply add the wireless network users to the authentication database instead of having to add the users to all the wireless access points with which the wireless devices might associate.

In a typical wireless authentication scenario, the WLC couples with the authentication database to allow the wireless association to proceed or fail. Authentication databases commonly used are LDAP and RADIUS, although under some scenarios the WLC can also store a small user database locally that can be used for authentication purposes.

Supporting Wired Network

The supporting wired network is the portion of the system that serves as a path between WLCs, APs, and wired call elements. Because the APs need or might need to communicate to the wired world, part of the wired network has to enable those communications. The supporting wired network consists of the switches, routers, and wired medium (WAN links and optical links) that work together to communicate with the various components that form the architecture for voice and video over WLAN.

Wireless Collaboration Endpoints

The wireless collaboration endpoints are the components of the architecture for voice and video over WLAN that users employ to communicate with each other. These endpoints can be voice-only or enabled for both voice and video. When end users employ the wireless communications endpoints to call a desired destination, the endpoints in turn forward the request to their associated call processing server. If the call is allowed, the endpoints process the voice or video, encode it, and send it to the receiving device or the next hop of processing. Typical Cisco wireless endpoints are wireless IP phones, voice and video software clients running on desktop computers, mobile smart phones connected through wireless media, and mobile collaboration enterprise tablets.

Wired Call Elements

Whether the wireless collaboration endpoints initiate a session between each other or with wired endpoints, wired call elements are involved in some way. Wired call elements (gateways and call processing entities) are the supporting infrastructure, with voice and video endpoints coupled to that infrastructure.

Wired call elements are needed typically to address two requirements:

- Call Control, page 3-65
- Media Termination, page 3-65

Call Control

Cisco wireless endpoints require a call control or call processing server to route calls efficiently and to provide a feature-rich experience for the end users. The call processing entity resides somewhere in the wired network, either in the LAN or across a WAN.

Call control for the Cisco wireless endpoints is achieved through a call control protocol, either SIP or SCCP.

Media Termination

Media termination on wired endpoints occurs when the end users of the wireless endpoints communicate with IP phones, PSTN users, or video endpoints. Voice gateways, IP phones, video terminals, PBX trunks, and transcoders all serve as termination points for media when a user communicates through them. This media termination occurs by means of coding and decoding of the voice or video session for the user communication.

High Availability for Voice and Video over WLAN

Providing high availability in collaboration solutions is a critical requirement for meeting the modern demands of continuous connectivity. Collaboration deployments designed for high availability increase reliability and up time. Using real-time applications such as voice or video over WLAN without high availability could have very adverse effects on the end user experience, including an inability to make voice or video calls.

Designing a solution for voice and video over WLAN with high availability requires focusing of the following main areas:

- Supporting Wired Network High Availability, page 3-66
- WLAN High Availability, page 3-66
- Call Processing High Availability, page 3-68

Supporting Wired Network High Availability

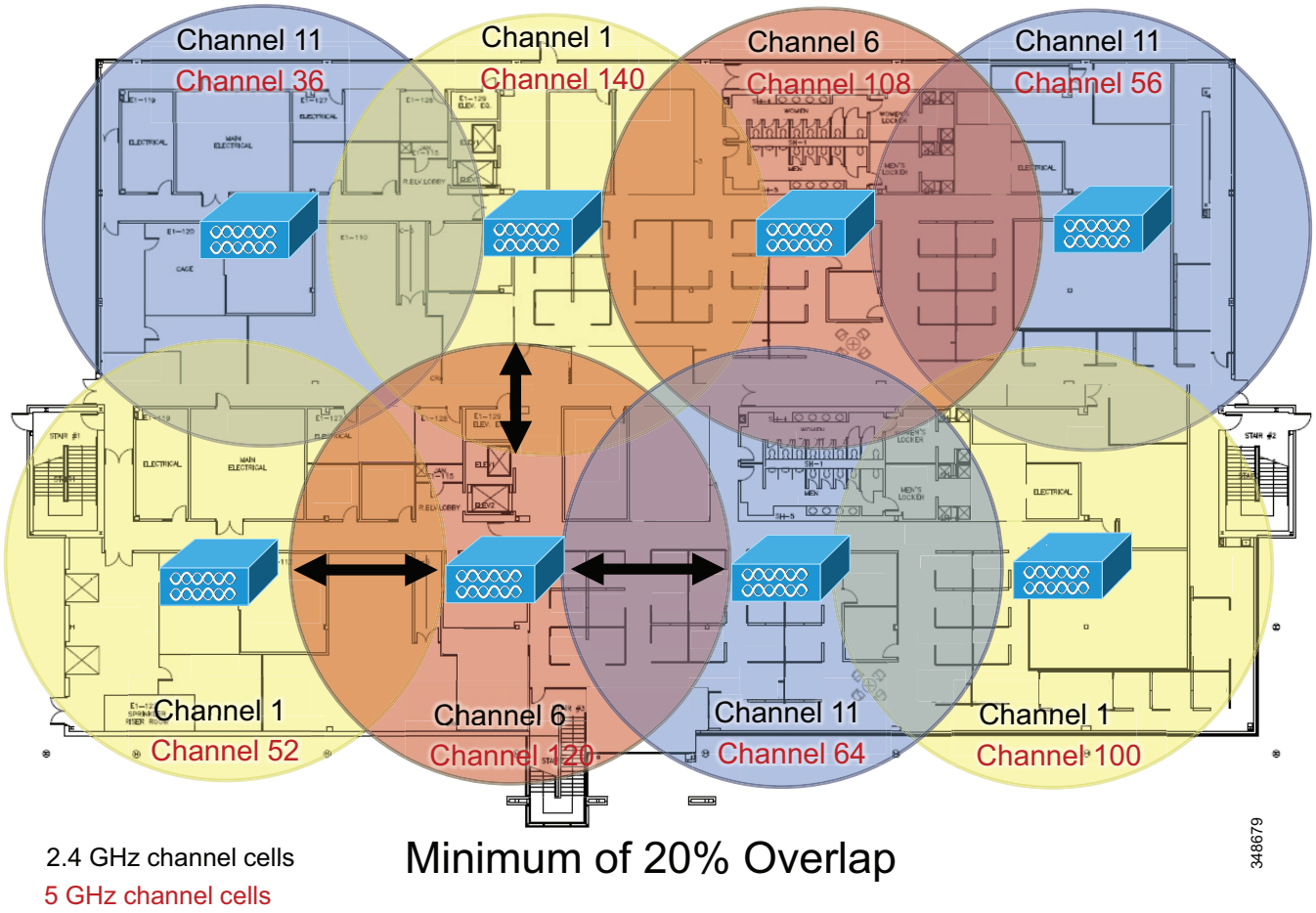
When deploying voice and video over WLAN, the same high-availability strategies used in wired networks can be applied to the wired components of the solution for voice and video over WLAN. For example, you can optimize layer convergence in the network to minimize disruption and take advantage of equal-cost redundant paths.

See LAN Design for High Availability, page 3-4, for further information about how to design highly available wired networks.

WLAN High Availability

A unique aspect of high availability for voice and video over WLAN is high availability of radio frequency (RF) coverage to provide Wi-Fi channel coverage that is not dependent upon a single WLAN radio. The Wi-Fi channel coverage is provided by the AP radios in the 2.4 GHz and 5 GHz frequency bands. The primary mechanism for providing RF high availability is cell boundary overlap. In general, a cell boundary overlap of 20% to 30% on non-adjacent channels is recommended to provide high availability in the wireless network. For mission-critical environments there should be at least two APs visible at the required signal level (-67 dBm or better). An overlap of 20% means that the RF cells of APs using non-adjacent channels overlap each other on 20% of their coverage area, while the remaining 80% of the coverage area is handled by a single AP. Figure 3-26 depicts a 20% overlap of AP non-adjacent channel cells to provide high availability. Furthermore, when determining the locations for installing the APs, avoid mounting them on reflective surfaces (such as metal, glass, and so forth), which could cause multi-path effects that result in signal distortion.

Figure 3-26 Non-Adjacent Channel Access Point Overlap



Careful deployment of APs and channel configuration within the wireless infrastructure are imperative for proper wireless network operation. For this reason, Cisco requires customers to conduct a complete and thorough site survey before deploying wireless networks in a production environment. The survey should include verifying non-overlapping channel configurations, Wi-Fi channel coverage, and required data and traffic rates; eliminating rogue APs; and identifying and mitigating the impact of potential interference sources.

Additionally, evaluate utilizing a 5 GHz frequency band, which is generally less crowded and thus usually less prone to interference. If Bluetooth is used then 5 GHz 802.11a is highly recommended. Similarly, the usage of Cisco CleanAir technology will increase the WLAN reliability by detecting radio frequency interference in real time and providing a self-healing and self-optimizing wireless network. For further information about Cisco CleanAir technology, refer to the product documentation available at

https://www.cisco.com/en/US/netsol/ns1070/index.html

For further information on how to provide high availability in a WLAN that supports rich media, refer to the *Real-Time Traffic over Wireless LAN Solution Reference Network Design Guide*, available at

https://www.cisco.com/en/US/docs/solutions/Enterprise/Mobility/RToWLAN/CCVP_BK_R7805F20_00_rtowlan-srnd.html

Call Processing High Availability

For information regarding call processing resiliency, see High Availability for Call Processing, page 9-13.

Capacity Planning for Voice and Video over WLAN

A crucial piece in planning for voice and video over WLAN is adequately sizing the solution for the desired call capacity. Capacity is defined as the number of simultaneous voice and video sessions over WLAN that can be supported in a given area. Capacity can vary depending upon the RF environment, the collaboration endpoint features, and the WLAN system features. For instance, a solution using Cisco Unified Wireless IP Phones 7925G on a WLAN that provides optimized WLAN services (such as the Cisco Unified Wireless Network) would have a maximum call capacity of 27 simultaneous sessions per channel at a data rate of 24 Mbps or higher for both 802.11a and 802.11g. On the other hand, a similar solution with a wireless device such as a tablet making video calls at 720p and a video rate of 2,500 kbps on a WLAN, where access points are configured as 802.11a/n with a data rate index of Modulation and Coding Scheme 7 in 40 MHz channels, would have a maximum capacity of 7 video calls (two bidirectional voice and video streams) per channel.

To achieve these capacities, there must be minimal wireless LAN background traffic and radio frequency (RF) utilization, and Bluetooth must be disabled in the devices. It is also important to understand that call capacities are established per non-overlapping channel because the limiting factor is the channel capacity and not the number of access points (APs).

The call capacity specified by the actual wireless endpoint should be used for deployment purposes because it is the supported capacity of that endpoint. For capacity information about the wireless endpoints, refer to the product documentation for your specific endpoint models:

https://www.cisco.com/c/en/us/products/collaboration-endpoints/product-listing.html

For further information about calculating call capacity in a WLAN, refer to the *Real-Time Traffic over Wireless LAN Solution Reference Network Design Guide*, available at

https://www.cisco.com/en/US/docs/solutions/Enterprise/Mobility/RToWLAN/CCVP_BK_R7805F20_00_rtowlan-srnd.html

Design Considerations for Voice and Video over WLAN

This section provides additional design considerations for deploying collaboration endpoints over WLAN solutions. WLAN configuration specifics can vary depending on the voice or video WLAN devices being used and the WLAN design. The following sections provide general guidelines and best practices for designing the WLAN infrastructure:

- VLANs, page 3-69
- Roaming, page 3-69
- Wireless Channels, page 3-69
- Wireless Interference and Multipath Distortion, page 3-70
- Multicast on the WLAN, page 3-71
- Wireless AP Configuration and Design, page 3-72
- Wireless LAN Controller Design Considerations, page 3-73
- WAN Quality of Service (QoS), page 3-37

VLANs

Just as with a wired LAN infrastructure, when deploying voice or video in a wireless LAN, you should enable at least two virtual LANs (VLANs) at the Access Layer. The Access Layer in a wireless LAN environment includes the access point (AP) and the first-hop access switch. On the AP and access switch, you should configure both a native VLAN for data traffic and a voice VLAN (under Cisco IOS) or Auxiliary VLAN (under CatOS) for voice traffic. This auxiliary voice VLAN should be separate from all the other wired voice VLANs in the network. However, when the wireless clients (for example, smart phones or software rich-media clients) do not support the concept of an auxiliary VLAN, alternative packet marking strategies (for example, packet classification per port) must be applied to segregate the important traffic such as voice and video and treat it with priority. When deploying a wireless infrastructure, Cisco also recommends configuring a separate management VLAN for the management of WLAN APs. This management VLAN should not have a WLAN appearance; that is, it should not have an associated service set identifier (SSID) and it should not be directly accessible from the WLAN.

Roaming

To improve the user experience, Cisco recommends designing the cell boundary distribution with a 20% to 30% overlap of non-adjacent channels to facilitate seamless roaming of the wireless client between access points. Furthermore, when devices roam at Layer 3, they move from one AP to another AP across native VLAN boundaries. When the WLAN infrastructure consists of autonomous APs, a Cisco Wireless LAN Controller allows the Cisco Unified Wireless endpoints to keep their IP addresses and roam at Layer 3 while still maintaining an active call. Seamless Layer 3 roaming occurs only when the client is roaming within the same mobility group. For details about the Cisco Wireless LAN Controller and Layer 3 roaming, refer to the product documentation available at

https://www.cisco.com/en/US/products/hw/wireless/index.html

Seamless Layer 3 roaming for clients across a lightweight access point infrastructure is accomplished by WLAN controllers that use dynamic interface tunneling. Cisco Wireless Unified Communications endpoints that roam across WLAN controllers and VLANs can keep their IP address when using the same SSID and therefore can maintain an active call.

Note In dual-band WLANs (those with 2.4 GHz and 5 GHz bands), it is possible to roam between 802.11b/g and 802.11a with the same SSID, provided the client is capable of supporting both bands. However, this can cause gaps in the voice path. If Cisco Unified Wireless IP Phones 7921 or 7925 are used, make sure that firmware version 1.3(4) or higher is installed on the phones to avoid these gaps; otherwise use only one band for voice. (The Cisco Unified Wireless IP Phone 7926 provides seamless inter-band roaming from its first firmware version.)

Wireless Channels

Wireless endpoints and APs communicate by means of radios on particular channels. When communicating on one channel, wireless endpoints typically are unaware of traffic and communication occurring on other non-overlapping channels.

Optimal channel configuration for 2.4 GHz 802.11b/g/n requires a minimum of five-channel separation between configured channels to prevent interference or overlap between channels. Non-overlapping channels have 22 MHz of separation. Channel 1 is 2.412 GHz, channel 6 is 2.437 GHz, and channel 11 is 2.462 GHz. In North America, with allowable channels of 1 to 11, channels 1, 6, and 11 are the three usable non-overlapping channels for APs and wireless endpoint devices. However, in Europe where the allowable channels are 1 to 13, multiple combinations of five-channel separation are possible. Multiple combinations of five-channel separation are also possible in Japan, where the allowable channels are 1 to 14.

Optimal channel configuration for 5 GHz 802.11a and 802.11n requires a minimum of one-channel separation to prevent interference or overlap between channels. In North America, there are 20 possible non-overlapping channels: 36, 40, 44, 48, 52, 56, 60, 64, 100, 104, 108, 112, 116, 132, 136, 140, 149, 153, 157, and 161. Europe and Japan allow 16 possible non-overlapping channels: 36, 40, 44, 48, 52, 56, 60, 64, 100, 104, 108, 112, 116, 132, 136, and 140. Because of the larger set of non-overlapping channels, 802.11a and 5 Ghz 802.11n allow for more densely deployed WLANs; however, Cisco recommends not enabling all channels but using a 12-channel design instead.

Note that the 802.11a and 802.11n bands (when using channels operating at 5.25 to 5.725 GHz, which are 15 of the 24 possible channels) do require support for Dynamic Frequency Selection (DFS) and Transmit Power Control (TPC) on some channels in order to avoid interference with radar (military, satellite, and weather). Regulations require that channels 52 to 64, 100 to 116, and 132 to 140 support DFS and TPC. TPC ensures that transmissions on these channels are not powerful enough to cause interference. DFC monitors channels for radar pulses and, when it detects a radar pulse, DFC stops transmission on the channel and switches to a new channel.

AP coverage should be deployed so that no (or minimal) overlap occurs between APs configured with the same channel. Same- channel overlap should typically occur at 19 dBm of separation. However, proper AP deployment and coverage on non-overlapping channels requires a minimum overlap of 20%. This amount of overlap ensures smooth roaming for wireless endpoints as they move between AP coverage cells. Overlap of less than 20% can result in slower roaming times and poor voice quality.

Deploying wireless devices in a multi-story building such as an office high-rise or hospital introduces a third dimension to wireless AP and channel coverage planning. Both the 2.4 GHz and 5.0 GHz wave forms of 802.11 can pass through floors and ceilings as well as walls. For this reason, not only is it important to consider overlapping cells or channels on the same floor, but it is also necessary to consider channel overlap between adjacent floors. With the 2.4 GHz wireless spectrum limited to only three usable non-overlapping channels, proper overlap design can be achieved only through careful three-dimensional planning.

Note Careful deployment of APs and channel configuration within the wireless infrastructure are imperative for proper wireless network operation. For this reason, Cisco requires that a complete and thorough site survey be conducted before deploying wireless networks in a production environment. The survey should include verifying non-overlapping channel configurations, AP coverage, and required data and traffic rates; eliminating rogue APs; and identifying and mitigating the impact of potential interference sources.

Wireless Interference and Multipath Distortion

Interference sources within a wireless environment can severely limit endpoint connectivity and channel coverage. In addition, objects and obstructions can cause signal reflection and multipath distortion. Multipath distortion occurs when traffic or signaling travels in more than one direction from the source to the destination. Typically, some of the traffic arrives at the destination before the rest of the traffic, which can result in delay and bit errors in some cases. You can reduce the effects of multipath distortion by eliminating or reducing interference sources and obstructions, and by using diversity antennas so that only a single antenna is receiving traffic at any one time. Interference sources should be identified during the site survey and, if possible, eliminated. At the very least, interference impact should be alleviated by proper AP placement and the use of location-appropriate directional or omni-directional diversity radio antennas.

Possible interference and multipath distortion sources include:

- Other APs on overlapping channels
- Other 2.4 GHz and 5 Ghz devices, such as 2.4 GHz cordless phones, personal wireless network devices, sulphur plasma lighting systems, microwave ovens, rogue APs, and other WLAN equipment that takes advantage of the license-free operation of the 2.4 GHz and 5 Ghz bands
- Metal equipment, structures, and other metal or reflective surfaces such as metal I-beams, filing cabinets, equipment racks, wire mesh or metallic walls, fire doors and fire walls, concrete, and heating and air conditioning ducts
- High-power electrical devices such as transformers, heavy-duty electric motors, refrigerators, elevators, and elevator equipment
- High-power electrical devices such as transformers, heavy-duty electric motors, refrigerators, elevators and elevator equipment, and any other power devices that could cause electromagnetic interference (EMI)

Because Bluetooth-enabled devices use the same 2.4 GHz radio band as 802.11b/g/n devices, it is possible that Bluetooth and 802.11b/g/n devices can interfere with each other, thus resulting in connectivity issues. Due to the potential for Bluetooth devices to interfere with and disrupt 802.11b/g/n WLAN voice and video devices (resulting in poor voice quality, de-registration, call setup delays, and/or reduce per-channel-cell call capacity), Cisco recommends, when possible, that you deploy all WLAN voice and video devices on the 5 GHz Wi-Fi band using 802.11a and/or 802.11n protocols. By deploying wireless clients on the 5 Ghz radio band, you can avoid interference caused by Bluetooth devices. Additionally, Cisco CleanAir technology is recommended within the wireless infrastructure because it enables real-time interference detection. For more information about Cisco CleanAir technology, refer to the product documentation available at

https://www.cisco.com/en/US/netsol/ns1070/index.html

Note 802.11n can operate on both the 2.4 GHz and 5 GHz bands; however, Cisco recommends using 5 GHz for Unified Communications.

Multicast on the WLAN

By design, multicast does not have the acknowledgement level of unicast. According to 802.11 specifications, the access point must buffer all multicast packets until the next Delivery Traffic Indicator Message (DTIM) period is met. The DTIM period is a multiple of the beacon period. If the beacon period is 100 ms (typical default) and the DTIM value is 2, then the access point must wait up to 200 ms before transmitting a single buffered multicast packet. The time period between beacons (as a product of the DTIM setting) is used by battery-powered devices to go into power save mode temporarily. This power save mode helps the device conserve battery power.

Multicast on WLAN presents a twofold problem in which administrators must weigh multicast traffic quality requirements against battery life requirements. First, delaying multicast packets will negatively affect multicast traffic quality, especially for applications that multicast real-time traffic such as voice and video. In order to limit the delay of multicast traffic, DTIM periods should typically be set to a value of 1 so that the amount of time multicast packets are buffered is low enough to eliminate any perceptible delay in multicast traffic delivery. However, when the DTIM period is set to a value of 1, the amount of time that battery-powered WLAN devices are able to go into power save mode is shortened, and therefore battery life is shortened. In order to conserve battery power and lengthen battery life, DTIM periods should typically be set to a value of 2 or more.

For WLAN networks with no multicast applications or traffic, the DTIM period should be set to a value of 2 or higher. For WLAN networks where multicast applications are present, the DTIM period should be set to a value of 2 with a 100 ms beacon period whenever possible; however, if multicast traffic quality

suffers or if unacceptable delay occurs, then the DTIM value should be lowered to 1. If the DTIM value is set to 1, administrators must keep in mind that battery life of battery-operated devices will be shortened significantly.

Before enabling multicast applications on the wireless network, Cisco recommends testing these applications to ensure that performance and behavior are acceptable.

For additional considerations with multicast traffic, see the chapter on Media Resources, page 7-1.

Wireless AP Configuration and Design

Proper AP selection, deployment, and configuration are essential to ensure that the wireless network handles voice traffic in a way that provides high-quality voice to the end users.

AP Selection

For recommends on deploying access points for wireless voice, refer to the documentation at https://www.cisco.com/en/US/products/ps5678/Products_Sub_Category_Home.html.

AP Deployment

The number of devices active with an AP affects the amount of time each device has access to the transport medium, the Wi-Fi channel. As the number of devices increases, the traffic contention increases. Associating more devices to the AP and the bandwidth of the medium can result in poor performance and slower response times for all the endpoint devices associated to the AP.

While there is no specific mechanism prior to Cisco Wireless LAN Controller release 7.2 to ensure that only a limited number of devices are associated to a single AP, system administrators can manage device-to-AP ratios by conducting periodic site surveys and analyzing user and device traffic patterns. If additional devices and users are added to the network in a particular area, additional site surveys should be conducted to determine whether additional APs are required to handle the number of endpoints that need to access the network.

Additionally, APs that support Cisco CleanAir technology should be considered because they provide the additional function of remote monitoring of the Wi-Fi channel.

AP Configuration

When deploying wireless voice, observe the following specific AP configuration requirements:

- Enable Address Resolution Protocol (ARP) caching.

  ARP caching is required on the AP because it enables the AP to answer ARP requests for the wireless endpoint devices without requiring the endpoint to leave power-save or idle mode. This feature results in extended battery life for the wireless endpoint devices.

- Enable Dynamic Transmit Power Control (DTPC) on the AP.

  This ensures that the transmit power of the AP matches the transmit power of the voice endpoints. Matching transmit power helps eliminate the possibility of one-way audio traffic. Voice endpoints adjust their transmit power based on the Limit Client Power (mW) setting of the AP to which they are associated.

- Assign a Service Set Identifier (SSID) to each VLAN configured on the AP.

  SSIDs enable endpoints to select the wireless VLAN they will use for sending and receiving traffic. These wireless VLANs and SSIDs map to wired VLANs. For voice endpoints, this mapping ensures priority queuing treatment and access to the voice VLAN on the wired network.

- Enable **QoS Element for Wireless Phones** on the AP.

  This feature ensures that the AP will provide QoS Basic Service Set (QBSS) information elements in beacons. The QBSS element provides an estimate of the channel utilization on the AP, and Cisco wireless voice devices use it to help make roaming decisions and to reject call attempts when loads are too high. The APs also provide 802.11e clear channel assessment (CCA) QBSS in beacons. The CCA-based QBSS values reflect true channel utilization.

- Configure two QoS policies on the AP, and apply them to the VLANs and interfaces.

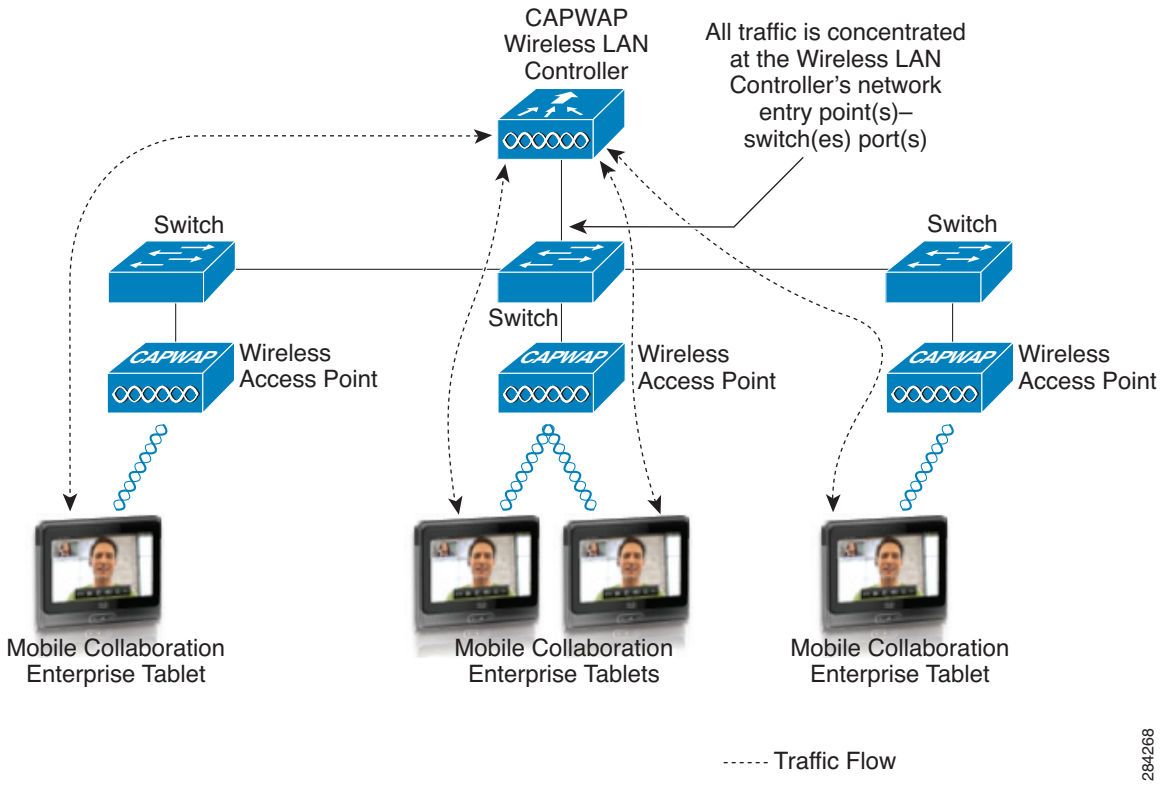
  To ensure that voice traffic is given priority queuing treatment, configure a voice policy and a data policy with default classifications for the respective VLANs. (See Interface Queuing, page 3-76, for more information).

Wireless LAN Controller Design Considerations

When designing a wireless network that will service voice or video, it is important to consider the role that the wireless LAN controller plays with regard to the voice and video media path if the access points used are not autonomous or stand alone. Because all wireless traffic is tunneled to its correspondent wireless LAN controller regardless of its point of origin and destination, it is critical to adequately size the network connectivity entry points of the wireless controllers. Figure 3-27 is a representation of this problem. If any mobile device tries to call another mobile device, the traffic has to be hairpinned in the wireless LAN controller and sent to the receiving device. This includes the scenario where both devices are associated to the same AP.

The switch ports where the wireless LAN controllers are connected should provide enough bandwidth coverage for the traffic generated by collaboration devices, whether they are video or voice endpoints and whether their traffic is control or media traffic.

Figure 3-27 Traffic Concentrated at the Wireless LAN Controller Network Entry Point



Additionally, the switch interface and switch platform egress buffer levels should match the maximum combined burst you plan to support in your wireless network.

Failure to select adequate buffer levels could lead to packet drops and severely affect the user experience of video over a wireless LAN, while lack of bandwidth coverage would cause packets to be queued and in extreme cases cause delayed packets

WLAN Quality of Service (QoS)

Just as QoS is necessary for the LAN and WAN wired network infrastructure in order to ensure high voice quality, QoS is also required for the wireless LAN infrastructure. Because of the bursty nature of data traffic and the fact that real-time traffic such as voice and video are sensitive to packet loss and delay, QoS tools are required to manage wireless LAN buffers, limit radio contention, and minimize packet loss, delay, and delay variation.

However, unlike most wired networks, wireless networks are a shared medium, and wireless endpoints do not have dedicated bandwidth for sending and receiving traffic. While wireless endpoints can mark traffic with 802.1p CoS, ToS, DSCP, and PHB, the shared nature of the wireless network means limited admission control and access to the network for these endpoints.

Wireless QoS involves the following main areas of configuration:

- Traffic Classification, page 3-75
- User Priority Mapping, page 3-75
- Interface Queuing, page 3-76
- Wireless Call Admission Control, page 3-77

Traffic Classification

As with the wired network infrastructure, it is important to classify or mark pertinent wireless traffic as close to the edge of the network as possible. Because traffic marking is an entrance criterion for queuing schemes throughout the wired and wireless network, marking should be done at the wireless endpoint device whenever possible. Marking or classification by wireless network devices should be identical to that for wired network devices, as indicated in Table 3-11.

In accordance with traffic classification guidelines for wired networks, the Cisco wireless endpoints mark voice media traffic or voice RTP traffic with DSCP 46 (or PHB EF), video media traffic or video RTP traffic with DSCP 34 (or PHB AF41), and call control signaling traffic (SCCP or SIP) with DSCP 24 (or PHB CS3). Once this traffic is marked, it can be given priority or better than best-effort treatment and queuing throughout the network. All wireless voice and video devices that are capable of marking traffic should do it in this manner. All other traffic on the wireless network should be marked as best-effort or with some intermediary classification as outlined in wired network marking guidelines. If the wireless voice or video devices are unable to do packet marking, alternate methods such as port-based marking should be implemented to provide priority to video and voice traffic.

User Priority Mapping

While 802.1p and Differentiated Services Code Point (DSCP) are the standards to set priorities on wired networks, 802.11e is the standard used for wireless networks. This is commonly referred as User Priority (UP), and it is important to map the UP to its appropriate DSCP value. Table 3-11 lists the values for collaboration traffic.

Table 3-11 QoS Traffic Classification

Traffic Type	DSCP (PHB)	802.1p UP	IEEE 802.11e UP
Voice	46 (EF)	5	6
Video	34 (AF41)	4	5
Voice and video control	24 (CS3)	3	4

For further information about 802.11e and its configuration, refer to your corresponding product documentation available at

https://www.cisco.com/en/US/products/ps6302/Products_Sub_Category_Home.html

Interface Queuing

Once traffic marking has occurred, it is necessary to enable the wired network APs and devices to provide QoS queuing so that voice and video traffic types are given separate queues to reduce the chances of this traffic being dropped or delayed as it traverses the wireless LAN. Queuing on the wireless network occurs in two directions, upstream and downstream. Upstream queuing concerns traffic traveling from the wireless endpoint up to the AP, and from the AP up to the wired network. Downstream queuing concerns traffic traveling from the wired network to the AP and down to the wireless endpoint.

For upstream queuing, devices that support Wi-Fi Multimedia (WMM) are able to take advantage of queueing mechanisms, including priority queueing.

As for downstream QoS, Cisco APs currently provide up to eight queues for downstream traffic being sent to wireless clients. The entrance criterion for these queues can be based on a number of factors, including DSCP, access control lists (ACLs), and VLAN. Although eight queues are available, Cisco recommends using only two queues when deploying wireless voice. All voice media and signaling traffic should be placed in the highest-priority queue, and all other traffic should be placed in the best-effort queue. This ensures the best possible queuing treatment for voice traffic.

In order to set up this two-queue configuration for autonomous APs, create two QoS policies on the AP. Name one policy **Voice**, and configure it with the class of service **Voice < 10 ms Latency (6)** as the Default Classification for all packets on the VLAN. Name the other policy **Data**, and configure it with the class of service **Best Effort (0)** as the Default Classification for all packets on the VLAN. Then assign the Data policy to the incoming and outgoing radio interface for the data VLAN(s), and assign the Voice policy to the incoming and outgoing radio interfaces for the voice VLAN(s). With the QoS policies applied at the VLAN level, the AP is not forced to examine every packet coming in or going out to determine the type of queuing the packet should receive.

For lightweight APs, the WLAN controller has built-in QoS profiles that can provide the same queuing policy. Voice VLAN or voice traffic is configured to use the **Platinum** policy, which sets priority queueing for the voice queue. Data VLAN or data traffic is configured to use the **Silver** policy, which sets best-effort queuing for the Data queue. These policies are then assigned to the incoming and outgoing radio interfaces based on the VLAN.

The above configurations ensure that all voice and video media and signaling are given priority queuing treatment in a downstream direction.

Note Because Wi-Fi Multimedia (WMM) access is based on Enhanced Distributed Channel Access (EDCA), it is important to assign the right priorities to the traffic to avoid Arbitration Inter-Frame Space (AIFS) alteration and delivery delay. For further information on Cisco Unified Wireless QoS, refer to the latest version of the *Enterprise Mobility Design Guide*, available at https://www.cisco.com/en/US/netsol/ns820/networking_solutions_design_guidances_list.html.

Wireless Call Admission Control

To avoid exceeding the capacity limit of a given AP channel, some form of call admission control is required. Cisco APs and wireless Unified Communications clients now use Traffic Specification (TSPEC) instead of QoS Basic Service Set (QBSS) for call admission control.

Wi-Fi Multimedia Traffic Specification (WMM TSPEC) is the QoS mechanism that enables WLAN clients to provide an indication of their bandwidth and QoS requirements so that APs can react to those requirements. When a client is preparing to make a call, it sends an Add Traffic Stream (ADDTS) message to the AP with which it is associated, indicating the TSPEC. The AP can then accept or reject the ADDTS request based on whether bandwidth and priority treatment are available. If the call is rejected, the client receives a Network Busy message. If the client is roaming, the TSPEC request is embedded in the re-association request message to the new AP as part of the association process, and the TSPEC response is embedded in the re-association response.

Alternatively, endpoints without WMM TSPEC support, but using SIP as call signaling, can be managed by the AP. Media snooping must be enabled for the service set identifier (SSID). The client's implementation of SIP must match that of the Wireless LAN Controller, including encryption and port numbers. For details about media snooping, refer to the *Cisco Wireless LAN Controller Configuration Guide*, available at

https://www.cisco.com/en/US/docs/wireless/controller/7.0/configuration/guide/c70wlan.html

Note Currently there is no call admission control support for video. The QoS Basic Service Set (QBSS) information element is sent by the AP only if **QoS Element for Wireless Phones** has been enable on the AP. (Refer to Wireless AP Configuration and Design, page 3-72.)

Cisco Collaboration Security

Revised: March 1, 2018

Securing the various components in a Cisco Collaboration Solution is necessary for protecting the integrity and confidentiality of voice and video calls.

This chapter presents security guidelines pertaining specifically to collaboration applications and the voice and video network. For more information on data network security, refer to the Cisco SAFE Blueprint documentation available at

https://www.cisco.com/c/en/us/solutions/enterprise/design-zone-security/index.html

Following the guidelines in this chapter does not guarantee a secure environment, nor will it prevent all penetration attacks on a network. You can achieve reasonable security by establishing a good security policy, following that security policy, staying up-to-date on the latest developments in the hacker and security communities, and maintaining and monitoring all systems with sound system administration practices.

This chapter addresses centralized and distributed call processing, including clustering over the WAN. This chapter assumes that all remote sites have a redundant link to the head-end or local call-processing backup in case of head-end failure. The interaction between Network Address Translation (NAT) and IP Telephony, for the most part, is not addressed here. This chapter also assumes that all networks are privately addressed and do not contain overlapping IP addresses.

What's New in This Chapter

Table 4-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 4-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Certificate management	Certificate Management, page 4-14	March 1, 2018
Encryption	Encryption, page 4-19	March 1, 2018
Security considerations for Cisco Unified Communications Manager (Unified CM)	Cisco Unified CM Security, page 4-21	March 1, 2018
Other miscellaneous updates	Various section of this chapter	March 1, 2018
Removed information on H.323	No longer in this document	March 1, 2018

General Security

This section covers general security features and practices that can be used to protect the voice data within a network.

Security Policy

Cisco Systems recommends creating a security policy associated with every network technology deployed within your enterprise. The security policy defines which data in your network is sensitive so that it can be protected properly when transported throughout the network. Having this security policy helps you define the security levels required for the types of data traffic that are on your network. Each type of data may or may not require its own security policy.

If no security policy exists for data on the company network, you should create one before enabling any of the security recommendations in this chapter. Without a security policy, it is difficult to ascertain whether the security that is enabled in a network is doing what it is designed to accomplish. Without a security policy, there is also no systematic way of enabling security for all the applications and types of data that run in a network.

Note While it is important to adhere to the security guidelines and recommendations presented in this chapter, they alone are not sufficient to constitute a security policy for your company. You must define a corporate security policy before implementing any security technology.

This chapter details the features and functionality of a Cisco Systems network that are available to protect the Unified Communications data on a network. It is up to the security policy to define which data to protect, how much protection is needed for that type of data, and which security techniques to use to provide that protection.

One of the more difficult issues with a security policy that includes voice and video traffic is combining the security policies that usually exist for both the data network and the traditional voice network. Ensure that all aspects of the integration of the media onto the network are secured at the correct level for your security policy or corporate environment.

The basis of a good security policy is defining how important your data is within the network. Once you have ranked the data according to its importance, you can decide how the security levels should be established for each type of data. You can then achieve the correct level of security by using both the network and application features.

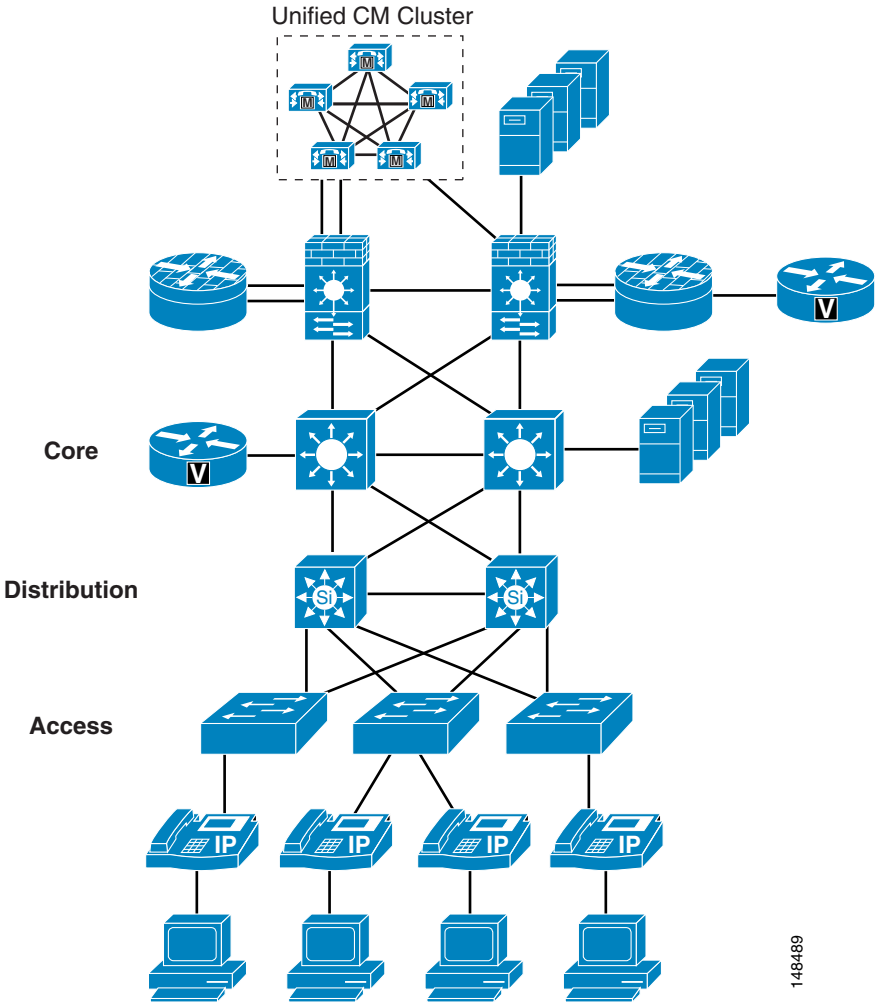
In summary, you can use the following process to define a security policy:

- Define the data that is on the network.
- Define the importance of that data.
- Apply security based on the importance of the data.

Security in Layers

This chapter starts with hardening the IP phone endpoints in a Cisco Unified Communications Solution and works its way through the network from the phone to the access switch, to the distribution layer, into the core, and then into the data center. (See Figure 4-1.) Cisco recommends building layer upon layer of security, starting at the access port into the network itself. This design approach gives a network architect the ability to place the devices where it is both physically and logically easy to deploy Cisco Unified Communications applications. But with this ease of deployment, the security complexity increases because the devices can be placed anywhere in a network as long as they have connectivity.

Figure 4-1 Layers of Security



Secure Infrastructure

As the IP Telephony data crosses a network, that data is only as safe and secure as the devices that are transporting the data. Depending on the security level that is defined in your security policy, the security of the network devices might have to be improved or they might already be secure enough for the transportation of IP Telephony traffic.

There are many best practices within a data network that, if used, will increase the entire security of your network. For example, instead of using Telnet (which sends passwords in clear text) to connect to any of the network devices, use Secure Shell (SSH, the secure form of Telnet) so that an attacker would not be able to see a password in clear text.

You should also use firewalls, access control lists, authentication services, and other Cisco security tools to help protect these devices from unauthorized access.

Physical Security

Just as a traditional PBX is usually locked in a secure environment, the IP network should be treated in a similar way. Each of the devices that carries media traffic is really part of an IP PBX, and normal general security practices should be used to control access to those devices. Once a user or attacker has physical access to one of the devices in a network, all kinds of problems could occur. Even if you have excellent password security and the user or attacker cannot get into the network device, that does not mean that they cannot cause havoc in a network by simply unplugging the device and stopping all traffic.

For more information on general security practices, refer to the documentation at the following locations:

- https://www.cisco.com/c/en/us/solutions/enterprise/design-zone-security/index.html
- https://www.cisco.com/c/en/us/products/security/service-listing.html

IP Addressing

IP addressing can be critical for controlling the data that flows in and out of the logically separated IP Telephony network. The more defined the IP addressing is within a network, the easier it becomes to control the devices on the network.

As stated in other sections of this document (see Campus Access Layer, page 3-4), you should use IP addressing based on RFC 1918. This method of addressing allows deployment of an IP Telephony system into a network without redoing the IP addressing of the network. Using RFC 1918 also allows for better control in the network because the IP addresses of the voice endpoints are well defined and easy to understand. If the voice and video endpoints are all addressed within a 10.x.x.x. network, access control lists (ACLs) and tracking of data to and from those devices are simplified.

If you have a well defined IP addressing plan for your voice deployments, it becomes easier to write ACLs for controlling the IP Telephony traffic and it also helps with firewall deployments.

Using RFC 1918 enables you easily to deploy one VLAN per switch, which is a best practice for campus design, and also enables you to keep the Voice VLAN free of any Spanning Tree Protocol (STP) loops.

If deployed correctly, route summarization could help to keep the routing table about the same as before the voice and video deployment, or just slightly larger.

IPv6 Addressing

The introduction of IPv6 addressing has extended the network address space and increased the options for privacy and security of endpoints. Though both IPv4 and IPv6 have similar security concerns, IPv6 provides some advantages. For example, one of the major benefits with IPv6 is the enormous size of the subnets, which discourages automated scanning and reconnaissance attacks.

When considering IPv6 as your IP addressing method, adhere to the best practices documented in the following campus and branch office design guides:

- *Deploying IPv6 in Campus Networks*

  https://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Campus/CampIPv6.html
- *Deploying IPv6 in Branch Networks*

  https://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Branch/BrchIPv6.html

Access Security

This section covers security features at the Access level that can be used to protect the voice and data within a network.

Voice and Video VLANs

Before the phone has its IP address, the phone determines which VLAN it should be in by means of the Cisco Discovery Protocol (CDP) negotiation that takes place between the phone and the switch. This negotiation allows the phone to send packets with 802.1q tags to the switch in a "voice VLAN" so that the voice data and all other data coming from the PC behind the phone are separated from each other at Layer 2. Voice VLANs are not required for the phones to operate, but they provide additional separation from other data on the network.

Voice VLANs can be assigned automatically from the switch to the phone, thus allowing for Layer 2 and Layer 3 separations between voice data and all other data on a network. A voice VLAN also allows for a different IP addressing scheme because the separate VLAN can have a separate IP scope at the Dynamic Host Configuration Protocol (DHCP) server.

Applications use CDP messaging from the phones to assist in locating phones during an emergency call. The location of the phone will be much more difficult to determine if CDP is not enabled on the access port to which that phone is attached.

There is a possibility that information could be gathered from the CDP messaging that would normally go to the phone, and that information could be used to discover some of the network. Not all devices that can be used for voice or video with Unified CM are able to use CDP to assist in discovering the voice VLAN.

Third-party endpoints do not support Cisco Discovery Protocol (CDP) or 802.1Q VLAN ID tagging. To allow device discovery when third-party devices are involved, use the Link Layer Discovery Protocol (LLDP). LLDP for Media Endpoint Devices (LLDP-MED) is an extension to LLDP that enhances support for voice endpoints. LLDP-MED defines how a switch port transitions from LLDP to LLDP-MED if it detects an LLDP-MED-capable endpoint. Support for both LLDP and LLDP-MED on IP phones and LAN switches depends on the firmware and device models. To determine if LLDP-MED is supported on particular phone or switch models, check the specific product documentation, release notes, and bulletins.

Note If an IP phone with LLDP-MED capability is connected to a Cisco Catalyst switch running an earlier Cisco IOS release that does not support LLDP, the switch might indicate that an extra device has been connected to the switch port. This can happen if the Cisco Catalyst switch is using Port Security to count the number of devices connected. The appearance of an LLDP packet might cause the port count to increase and cause the switch to disable the port. Verify that your Cisco Catalyst switch supports LLDP, or increase the port count to a minimum of three, before deploying Cisco IP Phones with firmware that supports LLDP-MED Link Layer protocol.

If your servers and/or clients are separated by a firewall, you might have to permit a wide range of TCP and UDP ports between these endpoints and the servers. Refer to the port utilization guide for each product. For example, for Unified CM refer to the latest version of the *System Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

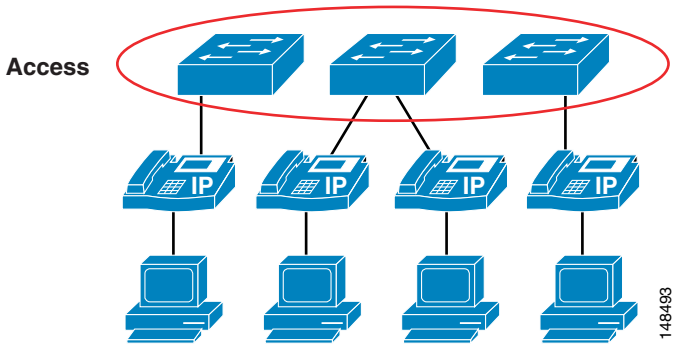
Gateways and servers are considered infrastructure devices, and they typically reside within the datacenter adjacent to the Unified CM servers. Clients, on the other hand, typically reside in the data VLAN.

Switch Port

There are many security features within a Cisco switch infrastructure that can be used to secure a data network. This section describes some of the features that can be used in Cisco Access Switches to protect the IP Telephony data within a network. (See Figure 4-2.) This section does not cover all of the security features available for all of the current Cisco switches, but it does list the most common security features used across many of the switches that Cisco manufactures. For additional information on the security features available on the particular Cisco gear deployed within your network, refer to the appropriate product documentation available at

https://www.cisco.com

Figure 4-2 A Typical Access Layer Design to Which the Phones Attach



Port Security: MAC CAM Flooding

A classic attack on a switched network is a MAC content-addressable memory (CAM) flooding attack. This type of attack floods the switch with so many MAC addresses that the switch does not know which port an end station or device is attached to. When the switch does not know which port a device is attached to, it broadcasts the traffic destined for that device to the entire VLAN. In this way, the attacker is able to see all traffic that is coming to all the users in a VLAN.

To disallow malicious MAC flooding attacks from hacker tools such as macof, limit the number of MAC addresses allowed to access individual ports based on the connectivity requirements for those ports. Malicious end-user stations can use macof to originate MAC flooding from random-source to random-destination MAC addresses, both directly connected to the switch port or through the IP phone. The macof tool is very aggressive and typically can fill a Cisco Catalyst switch content-addressable memory (CAM) table in less than ten seconds. The flooding of subsequent packets that remain unlearned because the CAM table is filled, is as disruptive and unsecure as packets on a shared Ethernet hub for the VLAN that is being attacked.

Either port security or dynamic port security can be used to inhibit a MAC flooding attack. A customer with no requirement to use port security as an authorization mechanism would want to use dynamic port security with the number of MAC addresses appropriate to the function attached to a particular port. For example, a port with only a workstation attached to it would want to limit the number of learned MAC addresses to one. A port with a Cisco Unified IP Phone and a workstation behind it would want to set the number of learned MAC addresses to two (one for the IP phone itself and one for the workstation behind the phone) if a workstation is going to plug into the PC port on the phone. This setting in the past has been three MAC addresses, used with the older way of configuring the port in trunk mode. If you use the multi-VLAN access mode of configuration for the phone port, this setting will be two MAC addresses, one for the phone and one for the PC plugged into the phone. If there will be no workstation on the PC port, then the number of MAC addresses on that port should be set to one. These configurations are for a multi-VLAN access port on a switch. The configuration could be different if the port is set to trunk mode (not the recommended deployment of an access port with a phone and PC).

Port Security: Prevent Port Access

Prevent all port access except from those devices designated by their MAC addresses to be on the port. This is a form of device-level security authorization. This requirement is used to authorize access to the network by using the single credential of the device's MAC address. By using port security (in its non-dynamic form), a network administrator would be required to associate MAC addresses statically for every port. However, with dynamic port security, network administrators can merely specify the number of MAC addresses they would like the switch to learn and, assuming the correct devices are the first devices to connect to the port, allow only those devices access to that port for some period of time.

The period of time can be determined by either a fixed timer or an inactivity timer (non-persistent access), or it can be permanently assigned. In the latter case, the MAC address learned will remain on the port even in the event of a reload or reboot of the switch.

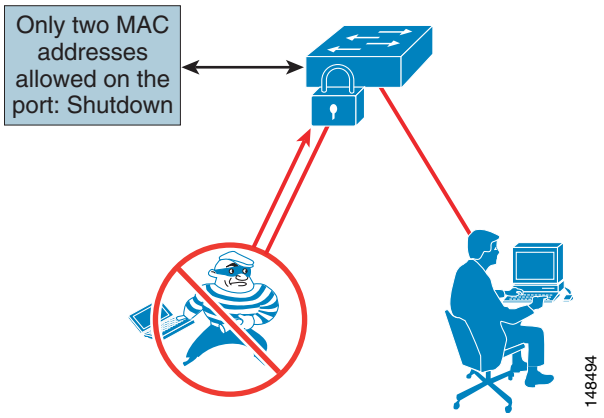
No provision is made for device mobility by static port security or persistent dynamic port security. Although it is not the primary requirement, MAC flooding attacks are implicitly prevented by port security configurations that aim to limit access to certain MAC addresses.

From a security perspective, there are better mechanisms for both authenticating and authorizing port access based on 802.1x rather than using MAC address authorization. MAC addresses alone can easily be spoofed or falsified by most operating systems.

Port Security: Prevent Rogue Network Extensions

Port security prevents an attacker from flooding the CAM table of a switch and from turning any VLAN into a hub that transmits all received traffic to all ports. It also prevents unapproved extensions of the network by adding hubs or switches into the network. Because it limits the number of MAC addresses to a port, port security can also be used as a mechanism to inhibit user extension to the IT-created network. For example, if a user plugs a wireless access point (AP) into a user-facing port or data port on a phone with port security defined for a single MAC address, the wireless AP itself would occupy that MAC address and not allow any devices behind it to access the network. (See Figure 4-3.) Generally, a configuration appropriate to stop MAC flooding is also appropriate to inhibit rogue access.

Figure 4-3 Limited Number of MAC Addresses Prevents Rogue Network Extensions



If the number of MAC addresses is not defined correctly, there is a possibility of denying access to the network or error-disabling the port and removing all devices from the network.

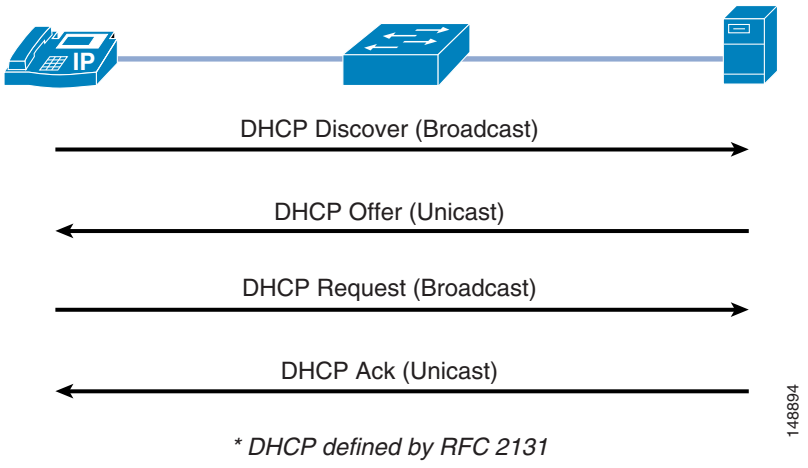
DHCP Snooping: Prevent Rogue DHCP Server Attacks

Dynamic Host Configuration Protocol (DHCP) Snooping prevents a non-approved DHCP or rogue DHCP server from handing out IP addresses on a network by blocking all replies to a DHCP request unless that port is allowed to reply. Because most phone deployments use DHCP to provide IP addresses to the phones, you should use the DHCP Snooping feature in the switches to secure DHCP messaging. Rogue DHCP servers can attempt to respond to the broadcast messages from a client to give out incorrect IP addresses, or they can attempt to confuse the client that is requesting an address.

When enabled, DHCP Snooping treats all ports in a VLAN as untrusted by default. An untrusted port is a user-facing port that should never make any reserved DHCP responses. If an untrusted DHCP-snooping port makes a DHCP server response, it will be blocked from responding. Therefore, rogue DHCP servers will be prevented from responding. However, legitimately attached DHCP servers or uplinks to legitimate servers must be trusted.

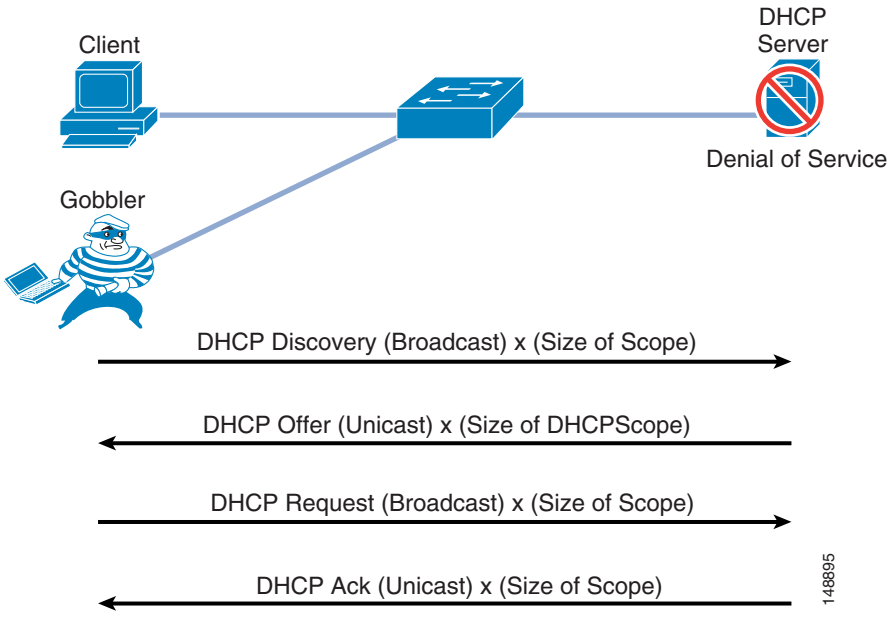
Figure 4-4 illustrates the normal operation of a network-attached device that requests an IP address from the DHCP server.

Figure 4-4 Normal Operation of a DHCP Request



However, an attacker can request not just a single IP address but all of the IP addresses that are available within a VLAN. (See Figure 4-5.) This means that there would be no addresses for a legitimate device trying to get on the network, and without an IP address the phone cannot connect to Unified CM.

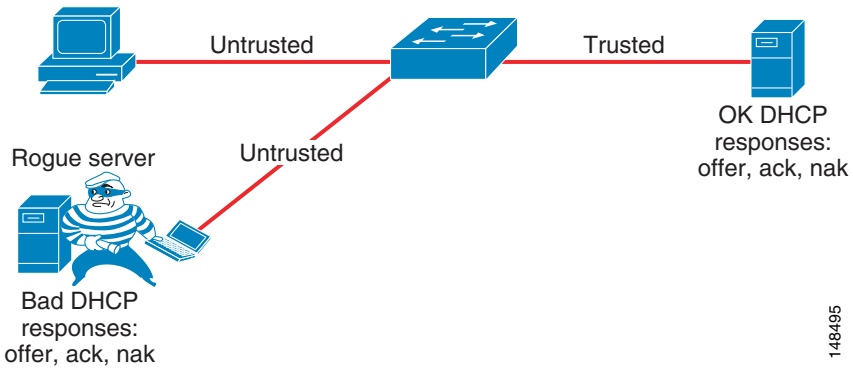
Figure 4-5 An Attacker Can Take All Available IP Addresses on the VLAN



DHCP Snooping: Prevent DHCP Starvation Attacks

DHCP address scope starvation attacks from tools such as Gobbler are used to create a DHCP denial-of-service (DoS) attack. Because the Gobbler tool makes DHCP requests from different random source MAC addresses, you can prevent it from starving a DHCP address space by using port security to limit the number of MAC addresses. (See Figure 4-6.) However, a more sophisticated DHCP starvation tool can make the DHCP requests from a single source MAC address and vary the DHCP payload information. With DHCP Snooping enabled, untrusted ports will make a comparison of the source MAC address to the DHCP payload information and fail the request if they do not match.

Figure 4-6 Using DHCP Snooping to Prevent DHCP Starvation Attacks



DHCP Snooping prevents any single device from capturing all the IP addresses in any given scope, but incorrect configurations of this feature can deny IP addresses to approved users.

DHCP Snooping: Binding Information

Another function of DHCP Snooping is to record the DHCP binding information for untrusted ports that successfully get IP addresses from the DHCP servers. The binding information is recorded in a table on the Cisco Catalyst switch. The DHCP binding table contains the IP address, MAC address, lease length, port, and VLAN information for each binding entry. The binding information from DHCP Snooping remains in effect for the length of the DHCP binding period set by the DHCP server (that is, the DHCP lease time). The DHCP binding information is used to create dynamic entries for Dynamic ARP Inspection (DAI) to limit ARP responses for only those addresses that are DHCP-bound. The DHCP binding information is also used by the IP source guard to limit sourcing of IP packets to only those addresses that are DHCP-bound.

There is a maximum limit to the number of binding table entries that each type of switch can store for DHCP Snooping. (Refer to the product documentation for your switch to determine this limit.) If you are concerned about the number of entries in your switch's binding table, you can reduce the lease time on the DHCP scope so that the entries in the binding table time-out sooner. The entries remain in the DHCP binding table until the lease runs out. In other words, the entries remain in the DHCP Snooping binding table as long at the DHCP server thinks the end station has that address. They are not removed from the port when the workstation or phone is unplugged.

If you have a Cisco Unified IP Phone plugged into a port and then move it to a different port, you might have two entries in the DHCP binding table with the same MAC and IP address on different ports. This behavior is considered normal operation.

Requirement for Dynamic ARP Inspection

Dynamic Address Resolution Protocol (ARP) Inspection (DAI) is a feature used on the switch to prevent Gratuitous ARP attacks on the devices plugged into the switch and on the router. Although it is similar to the Gratuitous ARP feature mentioned previously for the phones, Dynamic ARP protects all the devices on the LAN, and it is not just a phone feature.

In its most basic function, Address Resolution Protocol (ARP) enables a station to bind a MAC address to an IP address in an ARP cache, so that the two stations can communicate on a LAN segment. A station sends out an ARP request as a MAC broadcast. The station that owns the IP address in that request will give an ARP response (with its IP and MAC address) to the requesting station. The requesting station will cache the response in its ARP cache, which has a limited lifetime. The default ARP cache lifetime for Microsoft Windows is 2 minutes; for Linux, the default lifetime is 30 seconds; and for Cisco IP phones, the default lifetime is 40 minutes.

ARP also makes the provision for a function called Gratuitous ARP. Gratuitous ARP (GARP) is an unsolicited ARP reply. In its normal usage, it is sent as a MAC broadcast. All stations on a LAN segment that receive a GARP message will cache this unsolicited ARP reply, which acknowledges the sender as the owner of the IP address contained in the GARP message. Gratuitous ARP has a legitimate use for a station that needs to take over an address for another station on failure.

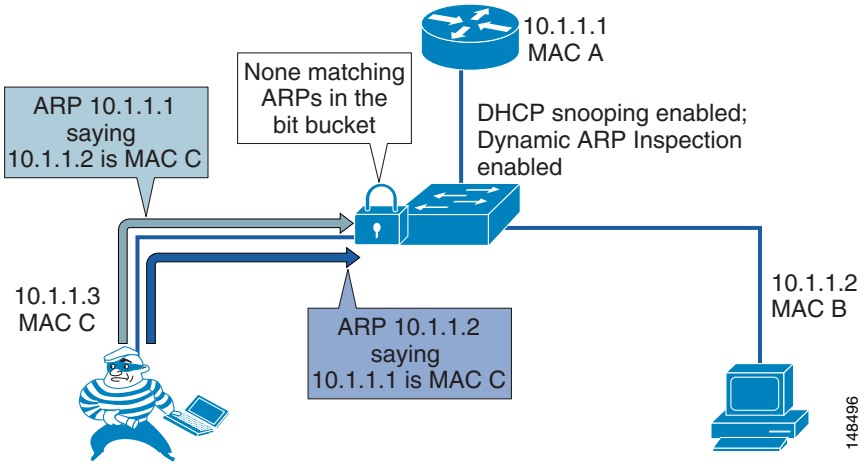
However, Gratuitous ARP can also be exploited by malicious programs that want to illegitimately take on the identity of another station. When a malicious station redirects traffic to itself from two other stations that were talking to each other, the hacker who sent the GARP messages becomes the man-in-the-middle. Hacker programs such as ettercap do this with precision by issuing "private" GARP messages to specific MAC addresses rather than broadcasting them. In this way, the victim of the attack does not see the GARP packet for its own address. Ettercap also keeps its ARP poisoning in effect by repeatedly sending the private GARP messages every 30 seconds.

Dynamic ARP Inspection (DAI) is used to inspect all ARP requests and replies (gratuitous or non-gratuitous) coming from untrusted (or user-facing) ports to ensure that they belong to the ARP owner. The ARP owner is the port that has a DHCP binding which matches the IP address contained in the ARP reply. ARP packets from a DAI trusted port are not inspected and are bridged to their respective VLANs.

Using DAI

Dynamic ARP Inspection (DAI) requires that a DHCP binding be present to legitimize ARP responses or Gratuitous ARP messages. If a host does not use DHCP to obtain its address, it must either be trusted or an ARP inspection access control list (ACL) must be created to map the host's IP and MAC address. (See Figure 4-7.) Like DHCP Snooping, DAI is enabled per VLAN, with all ports defined as untrusted by default. To leverage the binding information from DHCP Snooping, DAI requires that DHCP Snooping be enabled on the VLAN prior to enabling DAI. If DHCP Snooping is not enabled before you enable DAI, none of the devices in that VLAN will be able to use ARP to connect to any other device in their VLAN, including the default gateway. The result will be a self-imposed denial of service to any device in that VLAN.

Figure 4-7 Using DHCP Snooping and DAI to Block ARP Attacks



Because of the importance of the DHCP Snooping binding table to the use of DAI, it is important to back up the binding table. The DHCP Snooping binding table can be backed up to bootflash, File Transfer Protocol (FTP), Remote Copy Protocol (RCP), slot0, and Trivial File Transfer Protocol (TFTP). If the DHCP Snooping binding table is not backed up, the Cisco Unified IP Phones could lose contact with the default gateway during a switch reboot. For example, assume that the DHCP Snooping binding table is not backed up and that you are using Cisco Unified IP Phones with a power adapter instead of line power. When the switch comes back up after a reboot, there will be no DHCP Snooping binding table entry for the phone, and the phone will not be able to communicate with the default gateway unless the DHCP Snooping binding table is backed up and loads the old information before traffic starts to flow from the phone.

Incorrect configurations of this feature can deny network access to approved users. If a device has no entry in the DHCP Snooping binding table, then that device will not be able to use ARP to connect to the default gateway and therefore will not be able to send traffic. If you use static IP addresses, those addresses will have to be entered manually into the DHCP Snooping binding table. If you have devices that do not use DHCP again to obtain their IP addresses when a link goes down (some UNIX or Linux machines behave this way), then you must back up the DHCP Snooping binding table.

802.1X Port-Based Authentication

The 802.1X authentication feature can be used to identify and validate the device credentials of a Cisco Unified IP Phone before granting it access to the network. 802.1X is a MAC-layer protocol that interacts between an end device and a RADIUS server. It encapsulates the Extensible Authentication Protocol (EAP) over LAN, or EAPOL, to transport the authentication messages between the end devices and the switch. In the 802.1X authentication process, the Cisco Unified IP Phone acts as an 802.1X supplicant and initiates the request to access the network. The Cisco Catalyst Switch, acting as the authenticator, passes the request to the authentication server and then either allows or restricts the phone from accessing the network.

802.1X can also be used to authenticate the data devices attached to the Cisco Unified IP Phones. An EAPOL pass-through mechanism is used by the Cisco Unified IP Phones, allowing the locally attached PC to pass EAPOL messages to the 802.1X authenticator. The Cisco Catalyst Switch port needs to be configured in multiple-authentication mode to permit one device on the voice VLAN and multiple authenticated devices on the data VLAN.

Note Cisco recommends authenticating the IP phone before the attached data device is authenticated.

The multiple-authentication mode assigns authenticated devices to either a data or voice VLAN, depending on the attributes received from the authentication server when access is approved. The 802.1X port is divided into a data domain and a voice domain.

In multiple-authentication mode, a guest VLAN can be enabled on the 802.1x port. The switch assigns end clients to a guest VLAN when the authentication server does not receive a response to its EAPOL identity frame or when EAPOL packets are not sent by the client. This allows data devices attached to a Cisco IP Phone, that do not support 802.1X, to be connected to the network.

A voice VLAN must be configured for the IP phone when the switch port is in a multiple-host mode. The RADIUS server must be configured to send a Cisco Attribute-Value (AV) pair attribute with a value of **device-traffic-class=voice**. Without this value, the switch treats the IP phone as a data device.

Dynamic VLAN assignment from a RADIUS server is supported only for data devices.

When a data or a voice device is detected on a port, its MAC address is blocked until authorization succeeds. If the authorization fails, the MAC address remains blocked for 5 minutes.

When the 802.1x authentication is enabled on an access port on which a voice VLAN is configured and to which a Cisco IP Phone is already connected, the phone loses connectivity to the switch for up to 30 seconds.

Most Cisco IP Phones support authentication by means of X.509 certificates using the EAP-Transport Layer Security (EAP-TLS) or EAP-Flexible Authentication with Secure Tunneling (EAP-FAST) methods of authentication. Some of the older models that do not support either method can be authenticated using MAC Authentication Bypass (MAB), which enables a Cisco Catalyst Switch to check the MAC address of the connecting device as the method of authentication.

To determine support for the 802.1X feature configuration, refer to the product guides for the Cisco Unified IP Phones and the Cisco Catalyst Switches, available at https://www.cisco.com.

For configuration information, refer to the *IP Telephony for 802.1x Design Guide*, available at

https://www.cisco.com/en/US/docs/solutions/Enterprise/Security/TrustSec_1.99/IP_Tele/IP_Telephony_DIG.html

Certificate Management

Certificates are critical for establishing secure connections in a Cisco Collaboration deployment. They allow individuals, computers, and other services on the network to be authenticated. Implementing good certificate management provides a good level of protection while reducing complexity.

This section starts with a brief overview of the public key infrastructure (PKI), then it provides general guidance.

Brief PKI Overview

The public key infrastructure (PKI) provides a mechanism to secure communications and validate identities of communicating parties. Communications are made secure through encryption, and identities are validated through the use of public/private key pairs and digital identity certificates.

Public/Private Key Pair

A public and private key pair comprises two uniquely related cryptographic keys that are mathematically related. Whatever is encrypted with a public key may be decrypted only by its corresponding private key (which must be kept secret), and vice versa.

Certificates

A digital certificate is an electronic credential that is used to certify the identity of individuals, computers, and other services on a network. It is a wrapper around the public key. It provides information about the owner of the public key. It is used, for example, in a TLS handshake to authenticate the other party, or it can be used to digitally sign a file. Certificates deployed with Cisco Collaboration products are based on the X.509 standards. The certificates include the following information, among others:

- Public Key
- Common Name (CN)
- Organization Name (O)
- Issuer Name
- Validity period (Not before, not after)
- Extensions (optional) — For example, Subject Alternate Name (SAN)

A certificate can be self-signed or signed by a certificate authority (CA).

Certificate Validation During TLS Handshake

When a client initiates a TLS connection to a server, the server sends its certificate during the TLS handshake so that the client can authenticate the server. This happens, for example, when an administrator or end-user connects to the Unified CM pages or when the Jabber client starts and connects to the Unified CM UDS server, IM and Presence server, and Unity Connection server.

In some cases, the server also authenticates the client and requests the client to send its certificate. This is mutual authentication (mutual TLS, or MTLS) and it is used, for example, between Unified CM and Cisco endpoints in encrypted mode (configured with media and signaling encryption), with SIP trunks connecting two Unified CM clusters, or with SIP trunks connecting Unified CM to Cisco Unity Connection, a Cisco IOS Gateway, or Cisco Expressway (if TLS verify is configured on Expressway).

When a certificate is received, the verification consists of checking the following items:

- Identity — The subject or identity for which the certificate is issued must match the identity that the initiator of the session intended to reach. The hostname (FQDN) is checked against the Common Name (CN) or Subject Alternate Name (SAN) extension.
- Validity period — The current time and date must be within the certificate's validity range.
- Revocation status of the certificate
- Trust — The certificate must be trusted. A certificate is considered trusted if the signing (issuing) party is trusted. Trust with signing parties typically is established by importing the certificate of the signing party into a store of trusted certificates (trust store). See the section on CA-Signed Certificates Instead of Self-Signed Certificates, page 4-17, for more details.

General Guidance on Certificates

Some servers such as Cisco Unified CM and IM and Presence Service can have different certificates for the various system services. Some servers such as Cisco Expressway have only one certificate for the services they provide. Table 4-2 lists the server certificates for some of the most commonly deployed products. ECDSA certificates are not listed.

Table 4-2 Server Certificates for Common Cisco Collaboration Components

Service	Certificate	Description
Cisco Unified CM	Tomcat	Used for secure web connections. Also used for services such as LDAP, ILS, and LBM.
Cisco Unified CM	CallManager	Used for secure signaling by the CallManager service and for TFTP service to sign configuration files and ITL.
Cisco Unified CM	CAPF	Required by endpoints when connecting to the Certificate Authority Proxy Function (CAPF) service.
Cisco Unified CM	TVS	Required when connecting to the Trust Validation Service (TVS).
Cisco Unified CM	ITLRecovery	Certificate used as a trust anchor to recover the trust between the endpoints and Unified CM. Included in ITL and CTL files.
Cisco Unified CM	ipsec	For IPsec connections. IPsec can be enabled, but it is not covered in this document. The ipsec certificates are also used for the Cisco Unified CM Disaster Recovery System (DRS).
Cisco Unified CM	authz	Used for OAuth
IM and Presence Service	Tomcat	For SIP clients (Unified CM), web services, SOAP, and LDAP
IM and Presence Service	cup	For SIP Proxy, Presence Engine, and SIP federation
IM and Presence Service	cup-xmpp	For secure XMPP (IM)
IM and Presence Service	cup-xmpp-s2s	For secure XMPP federation
IM and Presence Service	ipsec	For IPsec. The ipsec certificates are also used for the Cisco Unified CM Disaster Recovery System (DRS).

Table 4-2 Server Certificates for Common Cisco Collaboration Components (continued)

Service	Certificate	Description
Cisco Unity Connection	Tomcat	Unity Connection web services certificate. Used for media and signaling encryption to the voicemail ports.
Cisco Unity Connection	ipsec	For IPsec
Cisco Expressway-C	Server	For all secure connections from/to Expressway-C
Cisco Expressway-E	Server	For all secure connections from/to Expressway-E
Cisco Meeting Server	Database client	For Cisco Meeting Servers with the Call Bridge service without a database, to connect securely to Cisco Meeting Server nodes with a database
Cisco Meeting Server	Certificates for Web Admin, Call Bridge, XMPP, Web Bridge, and database server	For simplicity, the same certificate could be used for all Cisco Meeting Server nodes and services.
Survivable Remote Site Telephony (SRST), Cisco IOS Gateway, and Cisco Unified Border Element	Cisco IOS certificate	With SRST, the SRST certificate is included in the configuration file of each endpoint.
Cisco Prime Collaboration Deployment	Tomcat	For web services
Cisco Prime Collaboration Provisioning	Provisioning	For provisioning web access

There are also other certificates that are based on ECDSA. See the section on RSA and ECDSA, page 4-16, for more details.

In general, the Cisco Collaboration servers are installed by default with a self-signed certificate. However, this is not the case for all products. For example, Cisco Meeting Server has no certificate installed by default.

Cisco Unified CM self-signed certificates are valid for 5 years, except the ITLRecovery certificate, which is valid for 20 years. The validity for this certificate is longer because it acts as a system-wide trust anchor.

RSA and ECDSA

Certificates for the Cisco Collaboration products are typically based on RSA (Rivest, Shamir, and Adelman) for public/private keys and digital signatures. Some products also support Elliptical Curve Digital Signature Algorithm (ECDSA) certificates, and they could be available with both self-signed and CA signed certificates. Both ECDSA and RSA certificates can coexist on the servers. At this time, because ECDSA on the endpoints is not supported and for simplicity and interoperability reasons, the general recommendation is use RSA certificates.

Note Encryption cipher suites based on ECDHE for the key exchange do not require certificates based on ECDSA; they can be negotiated with certificates based on RSA.

CA-Signed Certificates Instead of Self-Signed Certificates

When servers are installed, by default self-signed certificates are installed with most of the Cisco Collaboration products. To establish trust with a service based on a self-signed certificate, the server self-signed certificates must be imported into the trusted certificates store (or trust store) of all entities requiring secure connections to the service (clients). If not, with servers initiating the connections (for example, with Unified CM SIP trunks), the connection will fail. With Jabber and web browsers, users are prompted with warning messages and can accept the certificates, which then could be added to the trusted certificate store. This should be avoided because being prompted multiple times to accept a number of certificates during startup of the client is not a good user experience. Even more important is the fact that most users will not actually verify whether the presented certificate is correct by checking the certificate's fingerprint, and instead will just accept any certificate. This breaks the security concept of certificate-based authentication for secure session establishment.

Importing self-signed certificates can be handled if the set of communicating parties is small, but it becomes less practical for large numbers of communication peers. This is the main reason why we recommend replacing most default self-signed certificates with certificates that are signed by a CA. It simplifies certificate management. With CA-signed certificates, it is not necessary to import each server certificate in the client trust store; but instead, importing the root CA certificate to the client trust store is sufficient. On the server side, in general, the root CA certificate must also be imported to the server trust store; and if intermediate CA(s) are used, all the certificates in the certificate chain must also be imported to the server trust store. Using CA-signed certificates also allows for issuing new service certificates without having to update all client or server trusted certificate stores, as long as the signing CA's root certificate has already been added to the trusted certificate stores of all clients. A CA-signed certificate is also a requirement when using multi-server certificates.

As an example of the benefit of using a CA-signed certificate, if self-signed certificates are used with Jabber clients, the Unified CM Tomcat certificates (for UDS and for downloading the TFTP configuration file), the IM and Presence Tomcat and cup-xmpp certificates (for login and secure chat), and the Unity Connection Tomcat certificates (for visual voice mail) of all nodes would have to be imported into the trust store of each client running Jabber. However, with a CA-signed certificate, only the signing CA's root certificate needs to be imported.

In general, using a CA-signed certificate instead of a self-signed certificate is most beneficial for Tomcat certificates because they are widely used and are user-facing certificates. Using CA-signed certificates for the CallManager certificates is also beneficial because it allows the use of multi-server certificates (see the section on Multi-Server Certificates, page 4-18, for more details) and avoids importing all the CallManager certificates for all of the entities that connect to Unified CM subscribers via a SIP trunk.

However, it is not necessary to sign all of the certificates with a CA. Some certificates are used only for internal operations and are provided to the entity that needs them without any user intervention. For example, the Trust Verification Service (TVS) certificate is included in the Initial Trust List (ITL) file, and that file is automatically downloaded by the endpoints when they boot, restart, or reset. Similarly, the ITL recovery certificate is included in the Certificate Trust List (CTL) and Initial Trust List (ITL). Thus, there are no benefits to signing those certificates with an external CA. There are also no real benefits to signing the CAPF certificate by an external CA. It does not provide support for Certificate Authority Proxy Function (CAPF) certificate or endpoint Locally Significant Certificate (LSC) revocation. Also, when configuring phone VPN or 802.1x, importing the root CA certificate into the

ASA or Radius server's trust store is not sufficient. The CAPF certificate would still have to be imported because the endpoints do not send the certificate chain (and therefore do not send the CAPF certificate) during a TLS handshake.

Table 4-3 lists the certificates that Cisco recommends to be signed by a CA.

Table 4-3 Examples of Recommended Certificates to be Signed by a CA

Product	Certificate	Notes
Cisco Unified CM and IM and Presence Service	Tomcat	Used for various applications, including administrators and users accessing the web interface and Jabber accessing UDS and logging in.
Cisco Unified CM	CallManager	Used for various applications, including SIP trunks.
Cisco Unified CM	ipsec	Only if IPsec is used
IM and Presence Service	xmpp	
IM and Presence Service	xmpp-s2s	
Cisco Unity Connection	Tomcat	Used for various applications, including administrators and users accessing the web interface and Jabber accessing visual voicemail.
Cisco Expressway-C	Server	
Cisco Expressway-E	Server	Use a public CA.
Survivable Remote Site Telephony (SRST) and Cisco IOS Gateway	SRST and Cisco IOS Gateway	
Cisco Unified Border Element	Cisco IOS	In general, use an enterprise CA. If the SIP service provider supports encryption, use a public CA.
Cisco Meeting Server	Server	
Cisco Meeting Server	Database client	
Cisco TelePresence Management Suite (TMS)	Server	
Cisco Prime Collaboration Deployment	Tomcat	
Cisco Prime Collaboration Provisioning	Provisioning	

Multi-Server Certificates

To further simplify certificate management, a multi-server certificate can be used. Instead of having a certificate for each node, a single CA-signed certificate can be used across all the nodes in a cluster. A single corresponding private key is also used across all the nodes and is automatically propagated across the nodes. The servers covered in a multi-server certificate are listed in the Subject Alternative Names (SAN) extensions. Implementation of a multi-server certificate requires using a third-party CA in the deployment.

We recommend using multi-server certificates wherever available, as described in Table 4-4.

Table 4-4 Multi-Server Certificate Support

Product	Certificate	Notes
Unified CM and IM and Presence Service	Tomcat	Single Tomcat certificate across all the Unified CM and IM and Presence nodes in a cluster. Generate the Certificate Signing Request (CSR) and upload the CA-issued certificate on the Unified CM publisher node.
Unified CM	CallManager	
IM and Presence Service	xmpp	
IM and Presence Service	xmpp-s2s	
Unity Connection	Tomcat	

Note Wildcard certificates typically are not supported for the Cisco Collaboration products.

Public versus Private CA

Besides the requirement to use a public CA for the Expressway-E certificates, you could use either a public or enterprise CA (private or internal CA) to sign the various certificates of the Cisco Collaboration products in this document. The benefits of using a public CA include the fact that some clients and servers by default already trust major public CAs, and it is not necessary to establish trust between those devices and the public CA (import the CA certificate into the client trust store). With a public CA, your IT organization also does not have to install and maintain internal CA servers. But the major drawbacks of public CAs are the cost to issue certificates and restrictions that some public CAs might have.

The recommendation is to use an enterprise CA for all the certificates in Table 4-2 except for the Expressway-E certificates, which must be signed by a public CA, and except for the Cisco Unified Border Element certificate if the SIP service provider supports encryption. For more information, refer to the *Security* chapter in the latest version of the *Preferred Architecture for Cisco Collaboration Enterprise On-Premises Deployments CVD*, available at https://www.cisco.com/go/pa.

Encryption

With more services extending beyond the internal network, and with internal networks potentially subject to internal attacks, encryption and authentication are becoming increasingly critical.

Encryption protects against attacks such as eavesdropping, tampering, and session replay. If an unauthorized user is able to capture the traffic, he/she would not be able to decrypt the contents of the encrypted communication or modify it without knowing the encryption keys. Encryption also provides authentication through digital certificates when the encrypted communication is set up. The authentication can be one-way authentication; for example, for an administrator or end user using a web browser to access web services, where the client (browser) authenticates the web server but where the server does not authenticate the client (browser). Alternatively, the authentication can be two-way with Mutual TLS (MTLS), where the server also authenticates the client. MTLS is used, for example, with the signaling between endpoints and the Unified CM server they are registered to or with Unified CM SIP trunks.

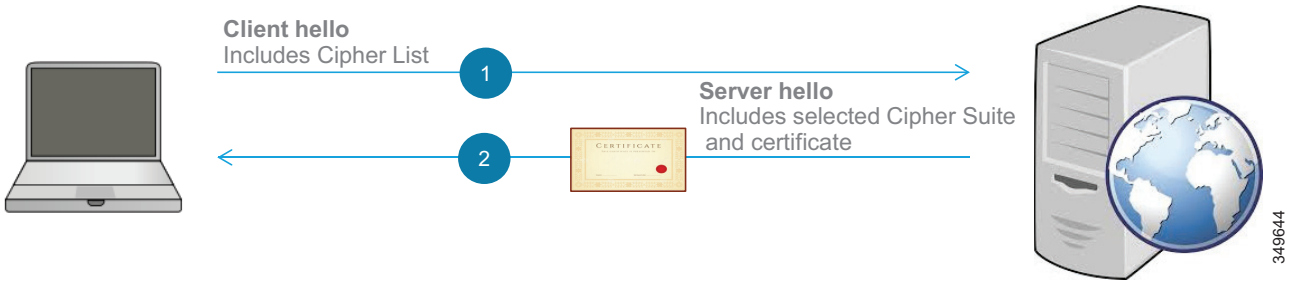
TLS Overview

Transport Layer Security (TLS) is a method for encrypting TCP traffic and is commonly used for web services traffic as well as SIP signaling. The following steps present an overview on how a TLS session is established:

1. A TLS connection is initiated by a TLS client, which connects to a TLS server. The client establishes a TCP connection with the server, sending first a Client Hello that contains a random number and its capabilities. These capabilities include the list of cipher suites the client supports.
2. The TLS server selects one of the cipher suites, typically taking into account the cipher suite preference of the client, and replies with a Server Hello. This message also includes another random number and the server certificate so that the client can authenticate it.

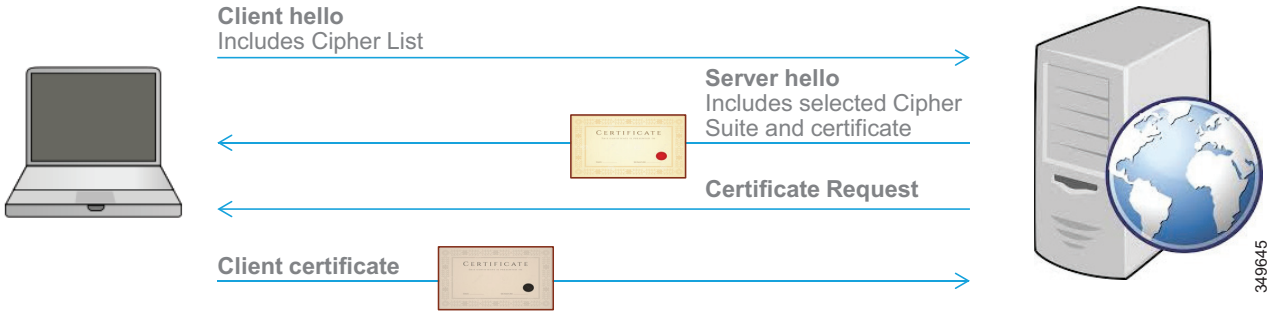
Figure 4-8 illustrates these two steps for establishing a TLS session. For simplicity, it does not include all the messages and possible variations in the TLS handshake. The server certificate could be sent in the Server Hello message or could be sent separately.

Figure 4-8 TLS Handshake



With Mutual TLS (MTLS), the server also authenticates the client. The server sends a CertificateRequest to the client, which in turn sends its client certificate. Figure 4-9 illustrates this flow at a high level.

Figure 4-9 MTLS Handshake



With RSA, the client encrypts the pre-master secret with the server's public key and sends it to the server. With Diffie-Hellman (DH) key agreement algorithms, the pre-master secret is not sent over the network; instead, the client and server exchange data (computed from random numbers and signed by the private key for authentication purposes) so that the client and the server can derive the pre-master secret on their own. DH combined with changing random numbers (Diffie-Hellman Ephemeral) allows for Perfect Forward Secrecy (PFS).

Then, the master secret is derived and session keys are computed from the master secret. From this point, the client and server stop using the public/private key pair (asymmetric encryption) and start using the shared session keys for encryption (symmetric encryption).

TLS is used to secure many of the communications links. For example, it is used to secure SIP or Skinny Client Control Protocol (SCCP) signaling.

In general, current Cisco Collaboration products support TLS version 1.2, With those products, TLS version 1.2 should always be negotiated, even if they also support TLS 1.0 and TLS 1.1. For security and/or compliance reasons, the administrator can choose to lock down the TLS version to 1.2, and therefore disable TLS 1.0 and TLS 1.1. This can be done with many Cisco Collaboration products. For the list of products that have this capability, refer to the *TLS 1.2 Compatibility Matrix for Cisco Collaboration Products*, available at

https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/uc_system/unified/communications/system/Compatibility/TLS/TLS1-2-Compatibility-Matrix.html

Also, for more details on TLS 1.2 support and disabling TLS 1.0 or 1.1, refer to the latest version of *TLS 1.2 for On-Premises Cisco Collaboration Deployments*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-system/products-configuration-examples-list.html

The Payment Card Industry Data Security Standard (PCI DSS) also has requirements on the version of TLS. For more details, refer to https://www.cisco.com/go/collabpci.

Media is secured using Secure RTP (SRTP), defined in IETF RFC 3711.

Cisco Unified CM Security

Hardened Platform

Cisco Unified CM and other Cisco Collaboration products based on the same platform run as a hardened appliance based on Linux OS, and they include the following default capabilities and restrictions:

- Root account is disabled.
- Third-party software installation is not allowed.
- Host-based intrusion protection (SELinux) and host-based firewall (IPTables) are installed and enabled by default.

  SELinux enforces policies that look at the behavior of the traffic to and from the server, and the way the applications are running on that server, to determine if everything is working correctly. If something is considered abnormal, then SELinux's access rules prevent that activity from happening.

  Connection rate limiting for DoS protection, and network shield protection for blocking specific ports, are configured using IPTables. The settings for the host-based firewall can be accessed using the Operating System Administration page of the Cisco Unified Communications server.

  SELinux cannot be disabled by an administrator, but it can be set to a permissive mode. It should be made permissive strictly for troubleshooting purposes. Disabling SELinux requires root access and can be done only by remote support from Cisco Technical Assistance Center (TAC).
- Complex password policy is applied to administrative accounts.
- Secure management interfaces (HTTPS, SSH, and SFTP) are enforced. Further, with the ability to assign users to access control groups and therefore to specific roles, administrators, end users, and application users can be given only the permissions they need.

- All installation packages are signed and include both the OS and the application.
- System audit logging is available, which is critical for determining what might have happened when issues arise.

Unified CM Mixed Mode for Media and Signaling Encryption

When Unified CM is first installed, it is in what we call "non-secure mode" even though most security features are actually available in this mode. For example, signed TFTP configuration file, encrypted TFTP configuration file, signed phone firmware, HTTPS access to web services, CAPF enrollment to install a Local Significant Certificate (LSC), SIP trunk encryption, Phone VPN, and 802.1x, are all possible by default with Unified CM in non-secure mode. The one security feature that is missing is media and signaling encryption for the endpoints. To enable it, Unified CM has to be configured in mixed mode, which requires Unified CM to be installed with the US Export Restricted version of the software (media and signaling encryption is not available with the Unrestricted version of Unified CM) and requires the Registration Token in Cisco Smart Software Licensing to be created with the export-controlled functionality.

There are two ways to enable mixed mode:

- Hardware USB eTokens

  This is the traditional way to enabled mixed mode. It requires a minimum of two Hardware USB eTokens (KEY-CCM-ADMIN-K9= or new KEY-CCM-ADMIN2-K9=). One eToken is used to sign the Certificate Trust List (CTL) file. The other eToken(s) provide redundancy in case the first eToken is lost or is not available anymore. To enable mixed mode, the CTL Client software must be installed onto a Microsoft Windows desktop. When this CTL client software is running, the USB eTokens will have to be inserted on the desktop. After mixed mode is configured, a CTL file is created for the Unified CM cluster, and the USB eTokens are removed and taken off-line.

- Tokenless (software eTokens)

  With this method, USB tokens and a Microsoft Windows desktop are not required. Mixed mode is enabled simply through a CLI command, **utils ctl set-cluster mixed-mode**. The CTL file is not signed by a hardware USB eToken but is signed by the Unified CM ITLRecovery private key.

  The tokenless method is recommended in general, and it provides the following benefits:

  – Enabling mixed mode and updating the CTL file is simpler. There is no need to acquire the USB eTokens, install the CTL client on a Microsoft Windows desktop, or run the CTL Client when enabling mixed mode or when updating the CTL file. Only one CLI command needs to be issued.

  – The key that signs the CTL file is longer with the tokenless method.

  – TLS 1.0 and 1.1 can be disabled on Unified CM with the tokenless option. With the USB hardware eTokens, the CTL client does not support TLS 1.1 or 1.2, so TLS 1.0 would have to be allowed.

  Note that, beginning with Cisco Unified CM 12.0, the tokenless CTL file (and ITL file) is signed by the ITLRecovery private key, so renewing the CallManager certificate is not a concern anymore and will not lead to a loss of trust between the endpoints and Unified CM.

Certificate Trust List (CTL) and Initial Trust List (ITL)

The Certificate Trust List (CTL) and Initial Trust List (ITL) are files that include Unified CM certificates. Those files are downloaded by Cisco endpoints when they boot, restart, or reset. These trust lists allow the endpoints to get the minimum set of Unified CM certificates to build the trust to Unified CM services. The ITL files are present in a Unified CM cluster, regardless of whether the Unified CM cluster is in non-secure mode or mixed mode. However, the CTL file is present and relevant only when Unified CM is in mixed mode.

The CTL and ITL files are signed by the System Administrator Security Token (SAST, see Table 4-5), and they contain a list of records. Each record contains a certificate, a certificate role or function, and pre-extracted certificate fields for easy look-up by the endpoints. Table 4-5 lists the certificate roles.

Table 4-5 Certificate Roles in CTL and ITL Files

Certificate Role	Certificates	Description
TFTP	CallManager	To authenticate Unified CM TFTP server. For example, used to verify TFTP configuration file signatures. Records with this certificate role are included in the ITL file when Unified CM is not in mixed mode.
CCM+TFTP	CallManager	To authenticate CallManager Service with encrypted signaling, and to authenticate the Unified CM TFTP server when verifying TFTP configuration file signatures. Records with this certificate role are included in the ITL and CTL files when Unified CM is in mixed mode.
System Administrator Security Token (SAST)	ITLRecovery and CallManager	To authenticate the SAST, which is the entity that signs the CTL, ITL, or TFTP configuration files. This type of record is included in the ITL and CTL files. The ITL and tokenless CTL files are signed by the ITL recovery key. The TFTP configuration files are signed by the TFTP servers' CallManager private keys.
Certificate Authority Proxy Function (CAPF)	CAPF	To authenticate CAPF service during secure communications with CAPF. A record with this certificate role is included in the ITL and CTL files if the CAPF service is activated on the Unified CM publisher.
Trusted Validated Service (TVS)	TVS	To authenticate TVS service when connecting to TVS. Present in the ITL file only.

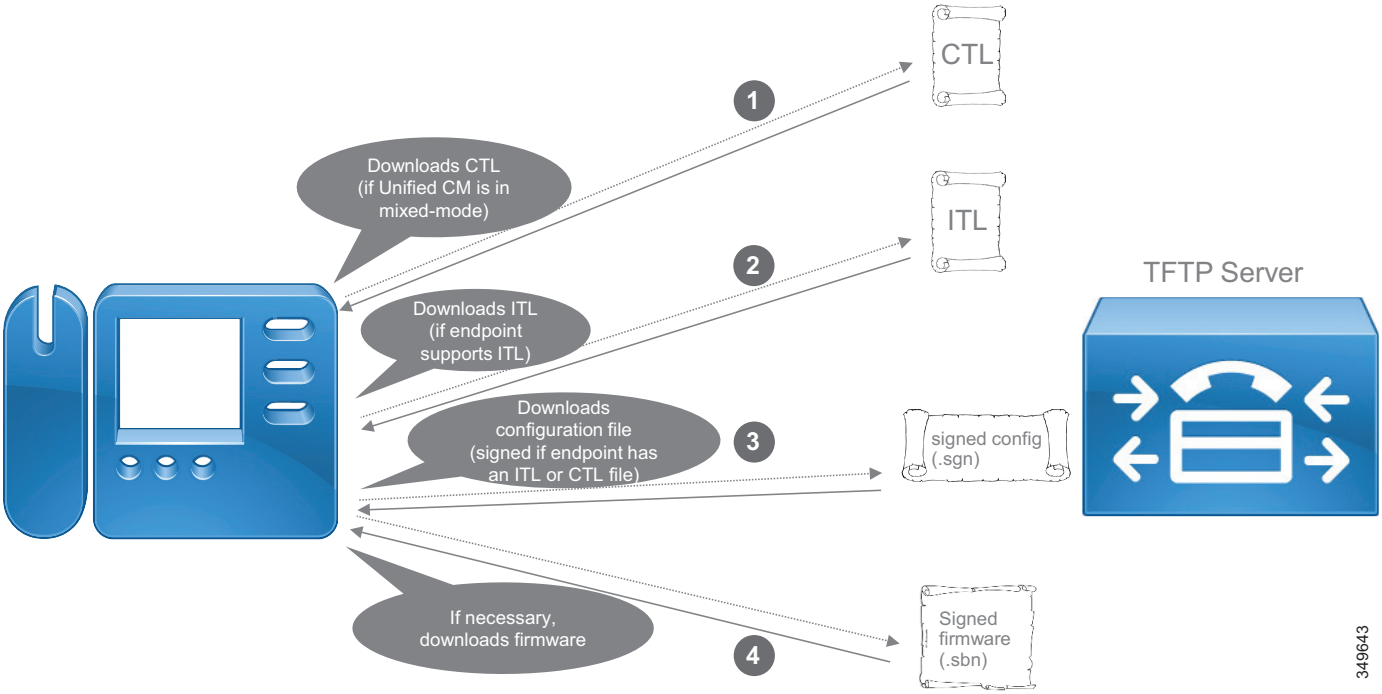
The ITL is signed by the ITLRecovery private key. Each Unified CM node running the TFTP service has its own ITL file that it provides to the endpoints.

The CTL file is signed by the private key of a System Administrator Security Token (SAST). With tokenless CTL, the SAST is the ITLRecovery private key. There is only one CTL file shared across the entire Unified CM cluster.

When endpoints boot or reset, before downloading their configuration file, they download the Certificate Trust List (CTL) from their TFTP server if Unified CM is in mixed mode. Then they download their TFTP server's Initial Trust List (ITL) if ITL is supported by the endpoint. Most Cisco endpoints support the ITL file, with some rare exceptions, Jabber being the main exception. If the endpoint is newly deployed and it is the first time the endpoint connects to Unified CM, it does not have an existing CTL or ITL file and therefore does not have a list of certificates it can use to validate the CTL or ITL signature. In that case, the endpoint simply accepts the CTL/ITL file in a one-time leap of faith and stores the certificates that are part of those files. Once the endpoint has a trusted list of certificates, it can use them to validate the signatures of subsequent CTL and ITL files that it downloads.

If an endpoint supports ITL or if Unified CM is in mixed mode (in which case a CTL file is downloaded by the endpoints), the endpoint possesses the CallManager certificate from the ITL/CTL file(s) and therefore requests a configuration file that is signed by the CallManager private key on the Unified CM TFTP server. If not (for example, as is the case with Jabber and when Unified CM is not in mixed mode), it requests a non-signed configuration file. After downloading its configuration file, the endpoint then verifies that it has the correct firmware. If it does not have the correct firmware, it downloads the relevant firmware and validates the signature of the firmware to ensure it was not tampered with. Figure 4-10 summarizes the files downloaded by the endpoints when they start up.

Figure 4-10 Files Downloaded by Endpoints During Startup



TFTP Configuration File Encryption

Without TFTP configuration file encryption, TFTP configuration files are available in plain text from any of the Unified CM TFTP servers. The type of information available in a TFTP configuration file includes, for example, phone firmware information and information on the Unified CM cluster. More importantly, if user names and passwords are provisioned in the Unified CM administration phone page, they are also saved in plain text in the TFTP configuration files. Therefore, the general recommendation is to enable TFTP configuration file encryption. This is especially important if user names, passwords, or sensitive information are configured in the Unified CM administration phone page.

However, with mobile and remote access (MRA) endpoints, if TFTP configuration file encryption is configured, the MRA endpoint must first be deployed on-premises and must register directly to Unified CM before being deployed in the Internet and connecting through MRA, even if it has an MIC. Moreover, with Jabber, if the endpoint is reset, it will not be able to get its encrypted configuration file and will not be able to register anymore until it is brought back inside the corporate network. For these reasons, it is simpler not to enable TFTP configuration file encryption for endpoints connecting through MRA and especially for Jabber endpoints connecting through MRA. However, ensure that no sensitive

information is configured for those endpoints. Therefore, we recommend that you disable TFTP configuration file encryption for those MRA endpoints (and do not provision passwords) but enable it for endpoints inside the corporate network. This is done by having a phone security profile with encrypted TFTP configuration enabled for the on-premises endpoints and a separate phone security profile with encrypted TFTP configuration disabled for the endpoints that connect through mobile and remote access (MRA).

Survivable Remote Site Telephony (SRST)

Cisco IOS SRST provides highly available call processing services for endpoints in locations remote from the Unified CM cluster. When endpoints cannot establish communications with the Unified CM call processing servers, they register to the local SRST router. SRST can be configured as non-secure or as secure. If SRST is configured as secure, endpoints configured in encrypted mode in Unified CM still have their media and signaling encrypted when registering to the SRST router.

After secure SRST is configured in Unified CM, Unified CM connects to the certificate provider service in the SRST router via a TLS connection on port 2445 by default and gets the certificate of the SRST router. This certificate is added to the endpoint TFTP configuration file so that when endpoints fail over to SRST, they can successfully authenticate the SRST router. The SRST router also authenticates the endpoints. Therefore, the certificate of the entity that signed the endpoint certificates must be imported to the SRST trust store. If LSC certificates are installed on the endpoints and if they are signed by CAPF, the CAPF certificate must be imported to the SRST trust store. If the phones are using their MIC certificates instead, the Cisco Manufacturing certificates must be imported to the SRST trust store.

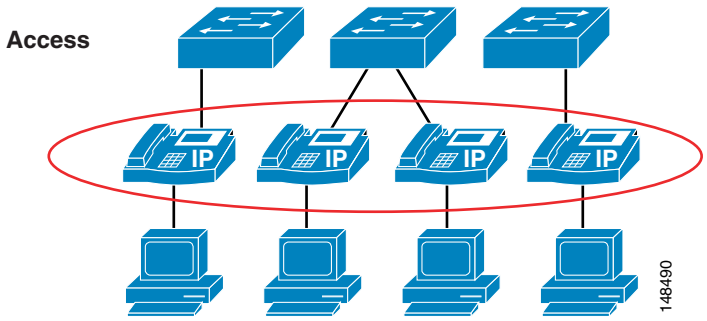
For more details, refer to the latest version of the *Cisco Unified SCCP and SIP SRST System Administrator Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-survivable-remote-site-telephony/products-installation-and-configuration-guides-list.html

Endpoint Security

Cisco Unified IP Phones contain built-in features to increase security on an IP Telephony network. These features can be enabled or disabled on a phone-by-phone basis to increase the security of an IP Telephony deployment. Depending on the placement of the phones, a security policy will help determine if these features need to be enabled and where they should be enabled. (See Figure 4-11.)

Figure 4-11 Security at the Phone Level



The following security considerations apply to IP phones:

- PC Port on the Phone, page 4-26
- PC Voice VLAN Access, page 4-26
- Web Access Through the Phone, page 4-27
- Settings Access, page 4-28
- Authentication and Encryption, page 4-29
- VPN Client for IP Phones, page 4-30

Before attempting to configure the security features on a phone, check the documentation at the following link to make sure the features are available on that particular phone model:

https://www.cisco.com/c/en/us/support/collaboration-endpoints/index.html

For more security information on Cisco Unified IP Phone 7800 and 8800 Series, refer to the *Cisco IP Phone 7800 and 8800 Series Security Overview White Paper*, available at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/unified-ip-phone-8800-series/white-paper-listing.html

PC Port on the Phone

The phone has the ability to turn on or turn off the port on the back of the phone, to which a PC would normally be connected. This feature can be used as a control point to access the network if that type of control is necessary.

Depending on the security policy and placement of the phones, the PC port on the back of any given phone might have to be disabled. Disabling this port would prevent a device from plugging into the back of the phone and getting network access through the phone itself. A phone in a common area such as a lobby would typically have its port disabled. Most companies would not want someone to get into the network on a non-controlled port because physical security is very weak in a lobby. Phones in a normal work area might also have their ports disabled if the security policy requires that no device should ever get access to the network through a phone PC port. Depending on the model of phone deployed, Cisco Unified Communications Manager (Unified CM) can disable the PC port on the back of the phone. Before attempting to enable this feature, check the documentation at the following link to verify that this features is supported on your particular model of Cisco Unified IP Phone:

https://www.cisco.com/c/en/us/support/collaboration-endpoints/index.html

PC Voice VLAN Access

Because there are two VLANs from the switch to the phone, the phone needs to protect the voice VLAN from any unwanted access. The phones can prevent unwanted access into the voice VLAN from the back of the phone. A feature called PC Voice VLAN Access prevents any access to the voice VLAN from the PC port on the back of the phone. When disabled, this feature does not allow the devices plugged into the PC port on the phone to "jump" VLANs and get onto the voice VLAN by sending 802.1q tagged information destined for the voice VLAN to the PC port on the back of the phone. The feature operates one of two ways, depending on the phone that is being configured. On the more advanced phones, the phone will block any traffic destined for the voice VLAN that is sent into the PC port on the back of the phone. In the example shown in Figure 4-12, if the PC tries to send any voice VLAN traffic (with an

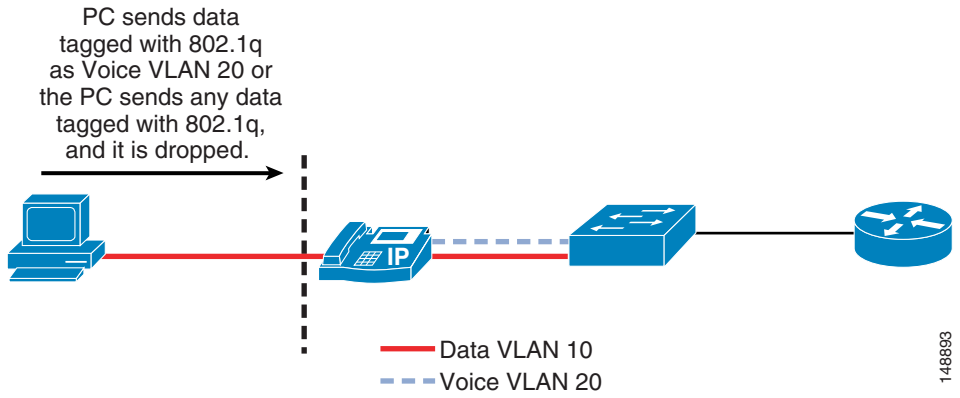
802.1q tag of 200 in this case) to the PC port on the phone, that traffic will be blocked. The other way this feature can operate is to block all traffic with an 802.1q tag (not just voice VLAN traffic) that comes into the PC port on the phone.

Currently, 802.1q tagging from an access port is not normally used. If that feature is a requirement for the PC plugged into the port on the phone, you should use a phone that allows 802.1q tagged packets to pass through the phone.

Before attempting to configure the PC Voice VLAN Access feature on a phone, check the documentation at the following link to make sure the feature is available on that particular phone model:

https://www.cisco.com/c/en/us/support/collaboration-endpoints/index.html

Figure 4-12 Blocking Traffic to the Voice VLAN from the Phone PC Port



Web Access Through the Phone

Each Cisco Unified IP Phone has a web server built into it to help with debugging and remote status of the phone for management purposes. The web server also enables the phones to receive applications pushed from Cisco Unified Communications Manager (Unified CM) to the phones. Access to this web server can be enabled or disabled on a phone by means of the Web Access feature in the Unified CM configuration. This setting can be global, or it could be enabled or disabled on a phone-by-phone basis.

If the web server is globally disable but it is needed to help with debugging, then the administrator for Unified CM will have to enable this feature on the phones. The ability to get to this web page can be controlled by an ACL in the network, leaving network operators with the capability to get to the web page when needed.

With the Web Access feature disabled, the phones will be unable to receive applications pushed to them from Unified CM.

Unified CM can be configured to use either HTTPS only or both HTTPS and HTTP for web traffic to and from the IP phones. However, if HTTPS only is configured, this does not by itself close port 80 on the IP phone's web server. It is preferable to use ACLs to restrict HTTP traffic, and configure Unified CM for HTTPS only.

Settings Access

Each Cisco Unified IP Phone has a network settings page that lists many of the network elements and detailed information that is needed for the phone to operate. This information could be used by an attacker to start a reconnaissance on the network with some of the information that is displayed on the phone's web page. For example, an attacker could look at the settings page to determine the default gateway, the TFTP server, and the Unified CM IP address. Each of these pieces of information could be used to gain access to the voice network or to attack a device in the voice network.

This access can be disabled on individual phones or by using bulk management to prevent end users or attackers from obtaining the additional information such as Unified CM IP address and TFTP server information. With access to the phone settings page disabled, end users lose the ability to change many of the settings on the phone that they would normally be able to control, such as speaker volume, contrast, and ring type. It might not be practical to use this security feature because of the limitations it places on end users with respect to the phone interface. The settings access can also be set as restricted, which prevents access to network configuration information but allows users to configure volume, ring tones, and so forth.

For more information on the phone settings page, refer to the latest version of the *Administration Guide for Cisco Unified Communications Manager and IM and Presence Service*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

Cisco TelePresence Endpoint Hardening

Cisco TelePresence endpoints have multiple configuration options for securing them against attacks. The security features vary among the different endpoints, and not all are enabled at default. These features include:

- Secure management over HTTPS and SSH
- Administrative passwords
- Device access
- Signaling and media encryption

Cisco TelePresence endpoints support management through Secure Shell (SSH) and Hyper-Text Transfer Protocol over Secure Sockets Layer (HTTPS). Access to the endpoints using HTTP, HTTPS, SSH, or Telnet can be configured in the Network Services setting on the endpoint itself.

The endpoints ship with default administrative passwords, and Cisco recommends changing the passwords at the time of installation. Access to management functions should be restricted to authorized users with administrative privileges. If the default administrative passwords are used, then the video stream can be viewed by anyone accessing the administrative page with the password.

The endpoints can be assigned to users who are given access based on defined roles and privileges. Passwords and PINs can be specified for those users to enable SSH or Telnet and web-based access. A credential management policy should be implemented to expire and change passwords periodically and to time-out logins when idle. This is necessary for limiting access to the devices to verified users.

Authentication and Encryption

Cisco Collaboration Solutions use Transport Layer Security (TLS) and Secure Real-time Transport Protocol (SRTP) for signaling and media encryption.

Transport Layer Security (TLS)

The Transport Layer Security (TLS) protocol is designed to provide authentication, data integrity, and confidentiality for communications between two applications. TLS operates in a client/server mode with one side acting as the "server" and the other side acting as the "client." TLS requires TCP as the reliable transport layer protocol to operate over.

Cisco Collaboration devices use TLS to secure SIP or Skinny Client Control Protocol (SCCP) signaling when connecting to Unified CM.

Secure Real-Time Transport Protocol (SRTP)

Secure RTP (SRTP), defined in IETF RFC 3711, details the methods of providing confidentiality and data integrity for both Real-time Transport Protocol (RTP) voice and video media, as well as their corresponding Real-time Transport Control Protocol (RTCP) streams. SRTP accomplishes this through the use of encryption and message authentication headers.

In SRTP, encryption applies only to the payload of the RTP packet. Message authentication, however, is applied to both the RTP header and the RTP payload. Because message authentication applies to the RTP sequence number within the header, SRTP indirectly provides protection against replay attacks as well. SRTP ciphers include Authentication Encryption with Associated Data (AEAD) with Advanced Encryption Standards (AES) 256 or 128 and with Secure Hash Algorithm (SHA) 2. SRTP cipher based on Hash-based Message Authentication Code with AES 128 and SHA-1 could also be negotiated if it is not disallowed in the configuration.

Voice and Video System

Unified CM can be configured to provide multiple levels of security to the phones within a voice system, if those phones support those features. This includes device authentication and media and signaling encryption using X.509 certificates. Depending on your security policy, phone placement, and phone support, the security can be configured to fit the needs of your company.

To enable security on the phones and in the Unified CM cluster, refer to the latest version of the *Security Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

When the Public Key Infrastructure (PKI) security features are properly configured in Unified CM, all supported phones will have the following capabilities:

- Integrity — Does not allow TFTP file manipulation with TFTP signed files but allows Transport Layer Security (TLS) signaling to the phones when enabled.
- Authentication — The image for the phone is authenticated from Unified CM to the phone, and the device (phone) is authenticated to Unified CM. All signaling messages between the phone and Unified CM are verified as being sent from the authorized device.
- Encryption — For supported devices, signaling and media can be encrypted to prevent eavesdropping. TFTP files can also be encrypted.
- Secure Real-time Transport Protocol (SRTP) — Is supported to Cisco IOS gateways and on phone-to-phone communications. Cisco Unity also supports SRTP for voicemail.

Unified CM supports authentication, integrity, and encryption for calls between two Cisco Unified IP Phones but not for all devices or phones. To determine if your device supports these features, refer to the documentation available at

https://www.cisco.com/c/en/us/support/collaboration-endpoints/index.html

Unified CM uses certificates for securing identities and enabling encryption. The certificates on the endpoints can be either Manufacturing Installed Certificates (MIC) or Locally Significant Certificates (LSC). MICs are pre-installed on most hardware endpoints, and LSCs are installed by Unified CM's Cisco Certificate Authority Proxy Function (CAPF). When MICs are used, the Cisco CA and the Cisco Manufacturing CA certificates act as the root certificates. When LSCs are generated for natively registered endpoints, they are most commonly signed by CAPF, and the CAPF certificate is the root certificate. It is also possible to use a third-party certificate authority (CA) to sign the LSC certificates (see the section on Third-Party CA Certificates, page 4-30).

Encryption on endpoints for signaling and media requires mixed mode on Unified CM. For more details on Unified CM mixed mode, see the section on Cisco Unified CM Security, page 4-21.

Cisco TelePresence Management Suite (TMS) provides TLS certificates to verify its identity when generating outbound connections.

Application layer protocol inspection and Application Layer Gateways (ALGs) that allow IP Telephony traffic to traverse firewalls and Network Address Translation (NAT) also do not work with signaling encryption. Not all gateways, phones, or conference are supported with encrypted media.

Encrypting media makes recording and monitoring of calls more difficult and expensive. It also makes troubleshooting VoIP problems more challenging.

Third-Party CA Certificates

By default, LSC certificates issued to the endpoints are signed by the CAPF service in Unified CM. However, third-party CA signed LSCs are also supported. Implementing such support involves importing the third-party CA certificate into the Unified CM trust store and configuring Unified CM's CAPF service to use the off-system CA as the certificate issuer for the endpoints. This method also requires heavy manual operations to handle the certificate signing requests (CSR) of all the phones, have them signed by a third-party CA, and import them back to the phones.

VPN Client for IP Phones

Cisco Unified IP Phones with an embedded VPN client provide a secure option for connecting phones outside the network to the Unified Communications solution in the enterprise. This functionality does not require an external VPN router at the remote location, and it provides a secure communications tunnel for Layer 3 and higher traffic over an untrusted network between the phone at the deployed location and the corporate network.

The VPN client in Cisco Unified IP Phones uses Cisco SSL VPN technology and can connect to both the Cisco ASA 5500 Series VPN head-end and the Cisco Integrated Services Routers with the Cisco IOS SSL VPN software feature. The voice traffic is carried in UDP and protected by Datagram Transport Layer Security (DTLS) protocol as part of the VPN tunnel. The integrated VPN tunnel applies only to voice and IP phone services. A PC connected to the PC port cannot use this tunnel and needs to establish its own VPN tunnel for any traffic from the PC.

For a phone with the embedded VPN client, you must first configure the phone with the VPN configuration parameters, including the VPN concentrator addresses, VPN concentrator credentials, user or phone ID, and credential policy. Because of the sensitivity of this information, the phone must be provisioned within the corporate network before the phone can attempt connecting over an untrusted network. Deploying the phone without first staging the phone in the corporate network is not supported.

The settings menu on the phone's user interface allows the user to enable or disable VPN tunnel establishment. When the VPN tunnel establishment is enabled, the phone starts to establish a VPN tunnel. The phone can be configured with up to three VPN concentrators to provide redundancy. The VPN client supports redirection from a VPN concentrator to other VPN concentrators as a load balancing mechanism.

For instructions on configuring the phones for the VPN client, refer to the latest version of the *Administration Guide for Cisco Unified Communications Manager and IM and Presence Service*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

Note For teleworkers and small offices or home offices (SOHOs), Cisco recommends deploying phone-based VPN, router-based VPN, or Mobile and Remote Access (MRA) with Cisco Expressway.

Quality of Service

Quality of Service (QoS) is a vital part of any security policy for an enterprise network. Even though most people think of QoS as setting the priority of traffic in a network, it also controls the amount of data that is allowed into the network. In the case of Cisco switches, that control point is at the port level when the data comes from the phone to the Ethernet switch. The more control applied at the edge of the network at the access port, the fewer problems will be encountered as the data aggregates in the network.

QoS can be used to control not only the priority of the traffic in the network but also the amount of traffic that can travel through any specific interface. Cisco Smartports templates have been created to assist in deploying voice QoS in a network at the access port level.

A rigorous QoS policy can control and prevent denial-of-service attacks in the network by throttling traffic rates.

As mentioned previously in the lobby phone example, Cisco recommends that you provide enough flow control of the traffic at the access port level to prevent any attacker from launching a denial-of-service (DoS) attack from that port in the lobby. The configuration for that example was not as aggressive as it could be because the QoS configuration allowed traffic sent to the port to exceed the maximum rate, but the traffic was remarked to the level of scavenger class. Given a more aggressive QoS policy, any amount of traffic that exceeded that maximum limit of the policy could just be dropped at the port, and that "unknown" traffic would never make it into the network. QoS should be enabled across the entire network to give the IP Telephony data high priority from end to end.

For more information on QoS, refer to the chapter on Network Infrastructure, page 3-1, and the QoS design guides available at

https://www.cisco.com/c/en/us/solutions/enterprise/design-zone-ipv6/design-guide-listing.html

Access Control Lists

This section covers access control lists (ACLs) and their uses in protecting voice data.

VLAN Access Control Lists

You can use VLAN access control lists (ACLs) to control data that flows on a network. Cisco switches have the capability of controlling Layers 2 to 4 within a VLAN ACL. Depending on the types of switches in a network, VLAN ACLs can be used to block traffic into and out of a particular VLAN. They can also be used to block intra-VLAN traffic to control what happens inside the VLAN between devices.

If you plan to deploy a VLAN ACL, you should verify which ports are needed to allow the phones to function with each application used in your IP Telephony network. Normally any VLAN ACL would be applied to the VLAN that the phones use. This would allow control at the access port, as close as possible to the devices that are plugged into that access port.

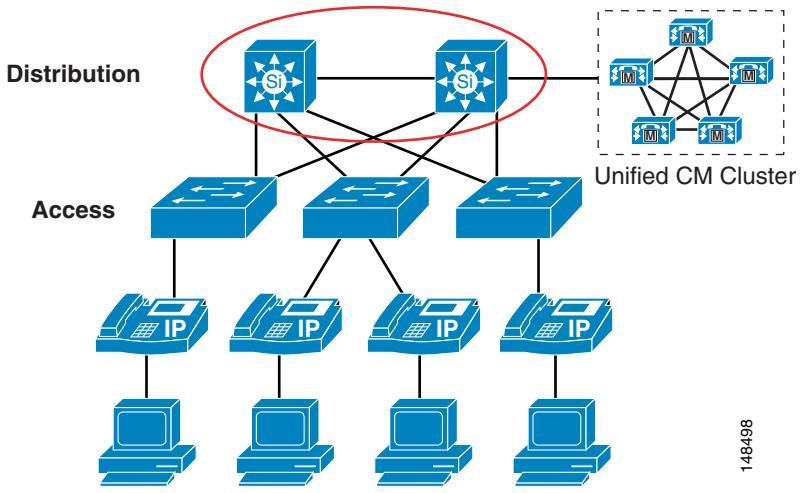
ACLs provide the ability to control the network traffic in and out of a VLAN as well as the ability to control the traffic within the VLAN.

VLAN ACLs are very difficult to deploy and manage at an access-port level that is highly mobile. Because of these management issues, care should be taken when deploying VLAN ACLs at the access port in the network.

Router Access Control Lists

As with VLAN ACLs, routers have the ability to process both inbound and outbound ACLs by port. The first Layer 3 device is the demarcation point between voice data and other types of data when using voice and data VLANs, where the two types of data are allowed to send traffic to each other. Unlike the VLAN ACLs, router ACLs are not deployed in every access device in your network. Rather, they are applied at the edge router, where all data is prepared for routing across the network. This is the perfect location to apply a Layer 3 ACL to control which areas the devices in each of the VLANs have the ability to access within a network. Layer 3 ACLs can be deployed across your entire network to protect devices from each other at points where the traffic converges. (See Figure 4-13.)

Figure 4-13 Router ACLs at Layer 3



There are many types of ACLs that can be deployed at Layer 3. For descriptions and examples of the most common types, refer to *Configuring Commonly Used IP ACLs*, available (with Cisco partner login required) at

https://www.cisco.com/c/en/us/support/docs/ip/access-lists/26448-ACLsamples.html

Depending on your security policy, the Layer 3 ACLs can be as simple as not allowing IP traffic from the non-voice VLANS to access the voice gateway in the network, or the ACLs can be detailed enough to control the individual ports and the time of the day that are used by other devices to communicate to IP Telephony devices. As the ACLs become more granular and detailed, any changes in port usage in a network could break not only voice but also other applications in the network.

If there are software phones in the network, if web access to the phone is allowed, or if you use the Attendant Console or other applications that need access to the voice VLAN subnets, the ACLs are much more difficult to deploy and control.

For IP phones restricted to specific subnets and limited to a voice VLAN, ACLs can be written to block all traffic (by IP address or IP range) to Unified CMs, voice gateways, phones, and any other voice application that is being used for voice-only services. This method simplifies the ACLs at Layer 3 compared to the ACLs at Layer 2 or VLAN ACLs.

Firewalls

Firewalls can be used in conjunction with ACLs to protect the voice servers and the voice gateways from devices that are not allowed to communicate with IP Telephony devices. Because of the dynamic nature of the ports used by IP Telephony, having a firewall does help to control opening up a large range of ports needed for IP Telephony communications. Given the complexities that firewalls introduce into a network design, you must take care in placing and configuring the firewalls and the devices around the firewalls to allow the traffic that is considered correct to pass while blocking the traffic that needs to be blocked.

IP Telephony networks have unique data flows. The phones use a client/server model for signaling for call setup, and Unified CM controls the phones through that signaling. The data flows for the IP Telephony RTP streams are more like a peer-to-peer network, and the phones or gateways talk directly

to each other via the RTP streams. If the signaling flows do not go through the firewall so that the firewall can inspect the signaling traffic, the RTP streams could be blocked because the firewall will not know which ports need to be opened to allow the RTP streams for a conversation.

A firewall placed in a correctly designed network can force all the data through that device, so capacities and performance need to be taken into account. Performance includes the amount of latency, which can be increased by a firewall if the firewall is under high load or even under attack. The general rule in an IP Telephony deployment is to keep the CPU usage of the firewalls to less than 60% for normal usage. If the CPU runs over 60%, it increases the chance of impacting IP phones, call setup, and registration. If the CPU usage stays at a sustained level above 60%, the registered IP phones will be affected, quality of calls in progress will degrade, and call setup for new calls will suffer. In the worst case, if the sustained CPU usage stays above 60%, phones will start to unregister. When this happens, they will attempt to re-register with Unified CM, thus increasing the load on the firewalls even more. If this were to happen, the effect would be a rolling blackout of phones unregistering and attempting to re-register with Unified CM. Until the CPU usage of the firewall decreases to under 60% sustained load, this rolling blackout would continue and most (if not all) of the phones would be affected. If you are currently using a Cisco firewall in your network, you should monitor the CPU usage carefully when adding IP Telephony traffic to your network so that you do not adversely affect that traffic.

There are many ways to deploy firewalls. This section concentrates on the Cisco Adaptive Security Appliance (ASA) in the active/standby mode in both routed and transparent scenarios. Each of the configurations in this section is in single-context mode within the voice sections of the firewall configurations.

All of the Cisco firewalls can run in either multiple-context or single-context mode. In single-context mode, the firewall is a single firewall that controls all traffic flowing through it. In multiple-context mode, the firewalls can be turned into many virtual firewalls. Each of these contexts or virtual firewalls have their own configurations and can be controlled by different groups or administrators. Each time a new context is added to a firewall, it will increase the load and memory requirements on that firewall. When you deploy a new context, make sure that the CPU requirements are met so that voice RTP streams are not adversely affected.

Adaptive Security Appliances have limited support for application inspection of IPv6 traffic for Unified Communications application servers and endpoints. Cisco recommends not using IPv6 for Unified Communications if ASAs are deployed in your network.

Note An ASA with No Payload Encryption module disables Unified Communications features.

A firewall provides a security control point in the network for applications that run over the network. A firewall also provides dynamic opening of ports for IP Telephony conversations if that traffic is running through the firewall.

Using its application inspection capability, the firewall can inspect the traffic that runs though it to determine if that traffic is really the type of traffic that the firewall is expecting. For example, does the HTTP traffic really look like HTTP traffic, or is it an attack? If it is an attack, then the firewall drops that packet and does not allow it to get to the HTTP server behind the firewall.

Not all IP Telephony application servers or applications are supported with firewall application layer protocol inspection. Some of these applications include Cisco Unity voicemail servers, Cisco Unified Attendant Console, Cisco Unified Contact Center Enterprise, and Cisco Unified Contact Center Express. ACLs can be written for these applications to allow traffic to flow through a firewall.

Note The timers for failover on the firewalls are set quite high by default. To keep from affecting voice RTP streams as they go through the firewall if there is a failover, Cisco recommends reducing those timer settings to less than one second. If this is done, and if there is a failover, the amount of time that the RTP streams could be affected will be less because the firewalls will fail-over quicker and there will be less impact on the RTP streams during the failover time.

When firewalls are placed between different Unified Communications components, the application inspection must be enabled for all protocols used for communications between the components. Application inspection can fail in call flow scenarios used by features such as Silent Monitoring by Unified Communications Manager, when the firewall is between the remote agent phones and the supervisor phones.

Unified Communications devices using TCP, such as Cisco Unified Communications Manager, support the TCP SACK option to speed up data transfer in case of packet loss. But not all firewalls support the TCP SACK option. In that case, TCP sessions established between Unified Communications devices through such a firewall will encounter problems if they attempt to use the TCP SACK option, and the TCP session might fail. Therefore, the firewalls should provide full support for the TCP SACK option. If support is not available, then the firewalls should be able to modify the TCP packets during the three-way handshake and to disable TCP SACK option support so that the endpoints will not attempt to use this option.

To determine if the applications running on your network are supported with the version of firewall in the network or if ACLs have to be written, refer to the appropriate application documentation available at

https://www.cisco.com

Routed ASA

The ASA firewall in routed mode acts as a router between connected networks, and each interface requires an IP address on a different subnet. In single-context mode, the routed firewall supports Open Shortest Path First (OSPF) and Routing Information Protocol (RIP) in passive mode. Multiple-context mode supports static routes only. ASA also supports Enhanced Interior Gateway Routing Protocol (EIGRP). Cisco recommends using the advanced routing capabilities of the upstream and downstream routers instead of relying on the security appliance for extensive routing needs. For more information on the routed mode, refer to the latest ASA configuration guides available at

https://www.cisco.com/c/en/us/support/security/adaptive-security-appliance-asa-software/products-installation-and-configuration-guides-list.html

The routed ASA firewall supports QoS, NAT, and VPN termination to the box, which are not supported in the transparent mode (see Transparent ASA, page 4-36). With the routed configuration, each interface on the ASA would have an IP address. In the transparent mode, there would be no IP address on the interfaces other then the IP address to manage the ASA remotely.

The limitations of this mode, when compared to the transparent mode, are that the device can be seen in the network and, because of that, it can be a point of attack. In addition, placing a routed ASA firewall in a network changes the network routing because some of the routing can be done by the firewall. IP addresses must also be available for all the interfaces on the firewall that are going to be use, so changing the IP addresses of the routers in the network might also be required. If a routing protocol or RSVP is to be allowed through the ASA firewall, then an ACL will have to be put on the inside (or most trusted) interface to allow that traffic to pass to the outside (or lesser trusted) interfaces. That ACL must also define all other traffic that will be allowed out of the most trusted interface.

Transparent ASA

The ASA firewall can be configured to be a Layer 2 firewall (also known as "bump in the wire" or "stealth firewall"). In this configuration, the firewall does not have an IP address (other than for management purposes), and all of the transactions are done at Layer 2 of the network. Even though the firewall acts as a bridge, Layer 3 traffic cannot pass through the security appliance unless you explicitly permit it with an extended access list. The only traffic allowed without an access list is Address Resolution Protocol (ARP) traffic.

This configuration has the advantage that an attacker cannot see the firewall because it is not doing any dynamic routing. Static routing is required to make the firewall work even in transparent mode.

This configuration also makes it easier to place the firewall into an existing network because routing does not have to change for the firewall. It also makes the firewall easier to manage and debug because it is not doing any routing within the firewall. Because the firewall is not processing routing requests, the performance of the firewall is usually somewhat higher with **inspect** commands and overall traffic than the same firewall model and software that is doing routing.

With transparent mode, if you are going to pass data for routing, you will also have to define the ACLs both inside and outside the firewall to allow traffic, unlike with the same firewall in routed mode. Cisco Discovery Protocol (CDP) traffic will not pass through the device even if it is defined. Each directly connected network must be on the same subnet. You cannot share interfaces between contexts; if you plan on running multiple-context mode, you will have to use additional interfaces. You must define all non-IP traffic, such as routing protocols, with an ACL to allow that traffic through the firewall. QoS is not supported in transparent mode. Multicast traffic can be allowed to go through the firewall with an extended ACL, but it is not a multicast device. In transparent mode, the firewall does not support VPN termination other than for the management interface.

If a routing protocol or RSVP is to be allowed through the ASA firewall, then an ACL will have to be put on the inside (or most trusted) interface to allow that traffic to pass to the outside (or lesser trusted) interfaces. That ACL must also define all other traffic that will be allowed out of the most trusted interface.

For more information on the transparent mode, refer to refer to the latest ASA configuration guides available at

https://www.cisco.com/c/en/us/support/security/adaptive-security-appliance-asa-software/products-installation-and-configuration-guides-list.html

Note Using NAT in transparent mode requires ASA version 8.0(2) or later. For more information, refer to the *Cisco ASA 5500 Series Release Notes* at https://www.cisco.com/c/en/us/support/security/asa-5500-series-next-generation-firewalls/products-release-notes-list.html.

Network Address Translation for Voice and Video

The Network Address Translation (NAT) device translates the private IP addresses inside the enterprise into public IP addresses visible on the public Internet. Endpoints inside the enterprise are internal endpoints, and endpoints in the public Internet are external endpoints.

When a device inside the enterprise connects out through the NAT, the NAT dynamically assigns a public IP address to the device. This public IP address is referred to as the *public mapped address* or the *reflexive transport address*. When the NAT forwards this packet to a device on the public Internet, the packet appears to come from its assigned public address. When external devices send packets back to the NAT at the public address, the NAT translates the IP addresses back to the internal private addresses and then forwards the packets to the internal network.

The NAT functionality is often part of the firewall and is therefore sometimes referred to as a NAT/FW. NATs map a large set of internal, private IP addresses into a smaller set of external, public IP addresses. The current public IPv4 address space is limited, and until IPv6 emerges as a ubiquitous protocol, most enterprises will have a limited number of IPv4 public addresses available. The NAT allows an enterprise with a large number of endpoints to make use of a small pool of public IP addresses. The NAT implements this functionality by dynamically mapping an internal IP address to an external IP address whenever an internal endpoint makes a connection out through the NAT. Each of these mappings is called a NAT binding.

The major complication in implementing NAT for voice and video devices occurs because the signaling protocols for voice and video include source addresses and ports in the protocol signaling messages. These source addresses provide the destination addresses that remote endpoints should use for return packets. However, internal endpoints use addresses from the private address space, and a NAT without an Application Layer Gateway (ALG) does not alter these internal addresses. When the remote endpoint receives a message, it cannot route packets to the private IP address in the message. Fixing this problem requires enabling an ALG (for example, a SIP or SCCP 'fixup') on the NAT device that can inspect the contents of the packet and implement address translation for the media IP addresses and port numbers encapsulated in the signaling messages.

A NAT ALG is similar to a firewall ALG, but a NAT ALG actually changes (maps) the addresses and ports in the signaling messages. The NAT ALG cannot inspect the contents of encrypted signaling messages.

Data Center

Within the data center, the security policy should define what security is needed for the IP Telephony applications servers. Because the Cisco Unified Communications servers are based on IP, the security that you would put on any other time-sensitive data within a data center could be applied to those servers as well.

If clustering over the WAN is being used between data centers, any additional security that is applied both within and between those data centers has to fit within the maximum round-trip time that is allowed between nodes in a cluster. In a multisite or redundant data center implementation that uses clustering over the WAN, if your current security policy for application servers requires securing the traffic between servers across data center firewalls, then Cisco recommends using IPSec tunnels for this traffic between the infrastructure security systems already deployed.

To design appropriate data center security for your data applications, Cisco recommends following the guidelines presented in the *Data Center Technology Design Guide*, available at

https://www.cisco.com/c/en/us/solutions/enterprise/data-center-designs-data-center-networking/index.html#~designs~tab-designs

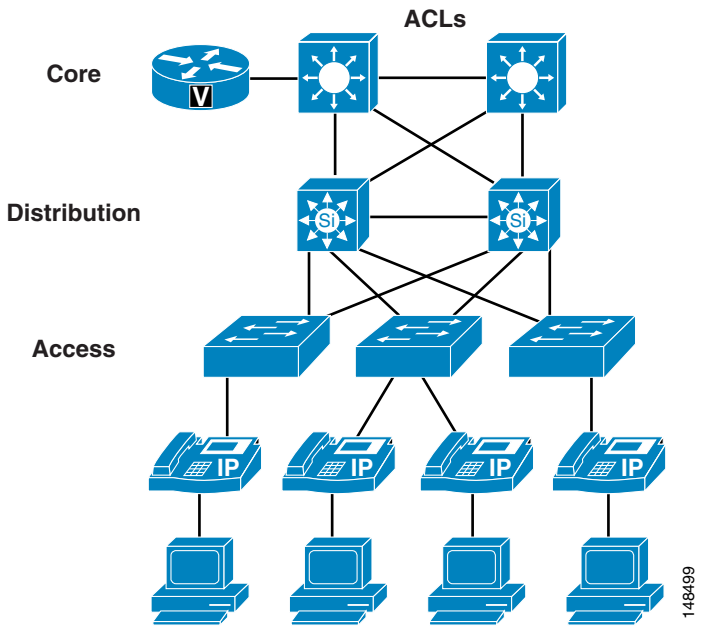
Gateways, Trunks, and Media Resources

Gateways and media resources are devices that convert an IP Telephony call into a PSTN call. When an outside call is placed, the gateway or media resource is one of the few places within an IP Telephony network to which all the voice RTP streams flow.

Because IP Telephony gateways and media resources can be placed almost anywhere in a network, securing an IP Telephony gateway or media resource might be considered more difficult than securing other devices, depending on your security policy. However, depending on which point trust is established in the network, the gateways and media resources can be quite easy to secure. Because of the way the gateways and media resources are controlled by Unified CM, if the path that the signaling takes to the gateway or media resource is in what is considered a secure section of the network, a simple ACL can be used to control signaling to and from the gateway or media resource. If the network is not considered secure between the gateways (or media resources) and where the Unified CMs are located (such as when a gateway is located at a remote branch), the infrastructure can be used to build IPSec tunnels to the gateways and media resources to protect the signaling. Most networks would most likely use a combination of the two approaches (ACL and IPSec) to secure those devices.

Because we use QoS at the edge of the network, if an attacker can get into the voice VLAN and determine where the gateways and media resources are, QoS at the port would limit how much data the attacker would be able to send to the gateway or media resource. (See Figure 4-14.)

Figure 4-14 Securing Gateways and Media Resources with IPSec, ACLs, and QoS



Some gateways and media resources support Secure RTP (SRTP) to the gateways and media resources from the phones, if the phone is enabled for SRTP. To determine if a gateway or media resource supports SRTP, refer to the appropriate product documentation at:

https://www.cisco.com

For more information on IPSec tunnels, refer to the *IPSec VPN WAN Design Overview*, available at:

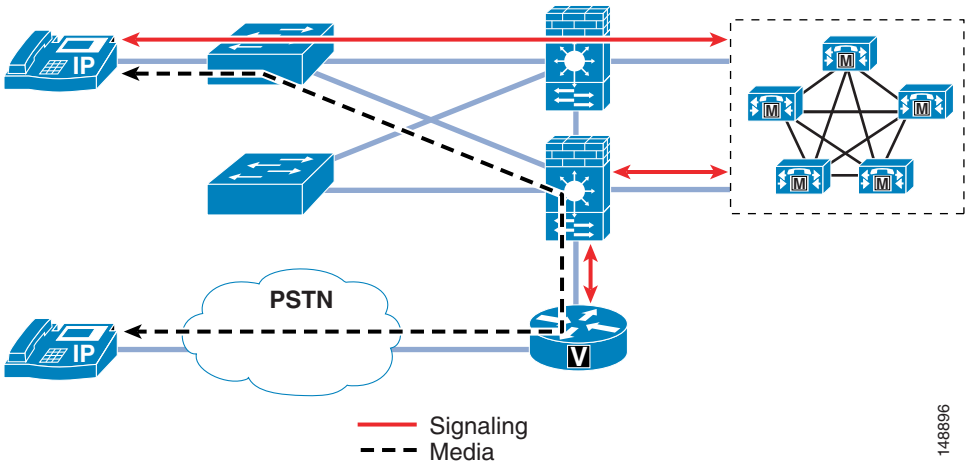
https://www.cisco.com/c/en/us/solutions/enterprise/design-zone-ipv6/design-guide-listing.html

Putting Firewalls Around Gateways

Some very interesting issues arise from placing firewalls between a phone making a call and the gateway to the PSTN network. Stateful firewalls look into the signaling messages between Unified CM, the gateway, and the phone, and they open a pinhole for the RTP streams to allow the call to take place. To do the same thing with a normal ACL, the entire port ranges used by the RTP streams would have to be open to the gateway.

There are two ways to deploy gateways within a network: behind a firewall and in front of a firewall. If you place the gateway behind a firewall, all the media from the phones that are using that gateway have to flow through the firewall, and additional CPU resources are required to run those streams through the firewall. In turn, the firewall adds control of those streams and protects the gateway from denial-of-service attacks. (See Figure 4-15.)

Figure 4-15 Gateway Placed Behind a Firewall



The second way to deploy the gateway is outside the firewall. Because the only type of data that is ever sent to the gateway from the phones is RTP streams, the access switch's QoS features control the amount of RTP traffic that can be sent to that gateway. The only thing that Unified CM sends to the gateway is the signaling to set up the call. If the gateway is put in an area of the network that is trusted, the only communication that has to be allowed between Unified CM and the gateway is that signaling. (See Figure 4-15.) This method of deployment decreases the load on the firewall because the RTP streams are not going through the firewall.

Unlike an ACL, most firewall configurations will open only the RTP stream port that Unified CM has told the phone and the gateway to use between those two devices as long as the signaling goes through the firewall. The firewall also has additional features for DoS attacks and Cisco Intrusion Prevention System (IPS) signatures to look at interesting traffic and determine if any attackers are doing something they should not be doing.

As stated in the section on Firewalls, page 4-33, when a firewall is looking at all the signaling and RTP streams from phones to a gateway, capacity could be an issue. Also, if data other than voice data is running through the firewall, CPU usage must be monitored to make sure that the firewall does not affect the calls that are running through the firewall.

Secure Audio and Video Conferencing

The Cisco IOS Enhanced Conference Bridges and Cisco Meeting Server provide secure conferencing. Implementing encrypted media and signaling between Unified CM, its endpoints, and the Cisco Meeting Server nodes, requires configuring the SIP trunk between the Unified CM server and the Cisco Meeting Server nodes as a secure SIP trunk. The SIP trunk configuration must also be set to allow SRTP. The Cisco Meeting Server certificates, or the CA certificate if CA-signed certificates are used, need to be uploaded to the Unified CM CallManager and Tomcat trust stores, and the certificate Common Name should be configured as the X.509 subject name on the SIP trunk profile. Likewise, the CallManager certificate, or the CA certificate if CA-signed certificates are used, should be uploaded to the Cisco Meeting Server trust store.

This configuration enables both secure signaling and HTTPS for management traffic between Cisco Unified CM and Cisco Meeting Server.

Unified CM Trunk Integration with Cisco Unified Border Element

Unified CM trunks provide an additional point of IP connectivity between the enterprise network and external networks. Additional security measures must be applied to these interconnects to mitigate threats inherent in data and IP telephony applications. Implementing a Cisco Unified Border Element between the Unified CM trunks and the external network provides for more flexible and secure interoperability options.

The Cisco Unified Border Element is a Cisco IOS software feature that provides voice application demarcation and security threat mitigation techniques applicable to both voice and data traffic. Cisco Unified Border Element can be configured in conjunction with Cisco IOS Firewall, Authentication, and VPN features on the same device to increase security for the Unified CM trunks integrated with service provider networks or other external networks. These Cisco IOS security features can serve as a defense against outside attacks and as a checkpoint for the internal traffic exiting to the service provider's network through the router. Infrastructure access control lists (ACLs) can also be used to prevent unauthorized access, DoS attacks, or distributed DoS (DDoS) attacks that originate from the service provider or a network connected to the service provider's network, as well as to prevent intrusions and data theft.

Cisco Unified Border Element is a back-to-back user agent (B2BUA) that provides the capability to hide network topology on signaling and media. It enables security and operational independence of the network and provides NAT service by substituting the Cisco Unified Border Element IP address on all traffic.

Cisco Unified Border Element can be used to re-mark DSCP QoS parameters on media and signaling packets between networks. This ensures that traffic adheres to QoS policies within the network.

Cisco IOS Firewall features, used in combination with Cisco Unified Border Element, provide Application Inspection and Control (AIC) to match signaling messages and manage traffic. This helps prevent SIP trunk DoS attacks and allows message filtering based on content and rate limiting.

Cisco Unified Border Element allows for SIP trunk registration. This capability is not available in Unified CM SIP trunks.

Cisco Unified Border Element can register the enterprise network's E.164 DID numbers to the service provider's SIP trunk on behalf of the endpoints behind it. If Cisco Unified Border Element is used to proxy the network's E.164 DID numbers, the status of the actual endpoint is not monitored. Therefore unregistered endpoints might still be seen as available.

Cisco Unified Border Element can connect RTP enterprise networks with SRTP over an external network. This allows secure communications without the need to deploy SRTP within the enterprise. It also supports RTP-SRTP interworking, but this is limited to a small number of codecs, including G.711 mulaw, G.711 alaw, G.729abr8, G.729ar8, G.729br8, and G.729r8.

Certain SIP service providers require SIP trunks to be registered before they allow call service. This ensures that calls originate only from well-known endpoints, thus making the service negotiation between the enterprise and the service provider more secure. Unified CM does not support registration on SIP trunks natively, but this support can be accomplished by using a Cisco Unified Border Element. The Cisco Unified Border Element registers to the service provider with the phone numbers of the enterprise on behalf of Cisco Unified Communications Manager.

For configuration and product details about Cisco Unified Border Element, refer to the documentation at:

- https://www.cisco.com/c/en/us/products/unified-communications/unified-border-element/index.html
- https://www.cisco.com/c/en/us/support/unified-communications/unified-border-element/products-installation-and-configuration-guides-list.html

Cisco Expressway in a DMZ

Cisco Expressway can establish video communication calls with devices outside the enterprise network and across the Internet with Expressway serving as the collaboration edge. Cisco Expressway-E must be placed outside the private network used by the Cisco Collaboration solution to allow external callers to access the device. It can be deployed either on the public Internet or in a demilitarized zone (DMZ). If Expressway-E is deployed in a DMZ, the firewall must be configured to allow traffic between the Internet and Expressway-E. Expressway-C is paired with Expressway-E and typically is deployed inside the data center. Together with Expressway-E, they facilitate firewall traversal for collaboration.

Expressway-C is a SIP Proxy and communications gateway for Cisco Unified CM. It is configured with a traversal client zone to communicate with Expressway-E to allow inbound and outbound calls to traverse the NAT device. Expressway-E is a SIP Proxy for devices that are located remotely (outside the enterprise network). It is configured with a public network domain name.

Cisco Expressway uses X.509 certificates for HTTPS, SIP TLS, and connections to systems such as Cisco Unified CM, LDAP, and syslog servers. It uses its list of trusted CA certificates, and it is preferable to use certificates signed by a third-party CA in this deployment. This simplifies the certificate configuration and exchange between Expressway-C, Expressway-E, and Unified CM, and it reduces management overhead. The Expressway-E certificate must be signed by a public third-party CA when physical endpoints are connected through mobile and remote access (MRA) because those endpoints use a built-in trust list of root CA certificates to validate the Expressway-E certificate.

Applications Servers

For a list of the Unified CM security features and how to enable them, refer to the latest version of the *Security Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

Before enabling any of the Unified CM security features, verify that they will satisfy the security requirements specified in your enterprise security policy for these types of devices in a network. For more information, refer to the *Cisco ASA 5500 Series Release Notes* at

https://www.cisco.com/c/en/us/support/security/asa-5500-series-next-generation-firewalls/products-release-notes-list.html

Single Sign-On

The Single Sign-On (SSO) feature allows for stronger authentication and a better user experience. For information on SSO, SAML authentication, and OAUTH protocols that are used with Cisco Collaboration solutions, refer to the chapter on Directory Integration and Identity Management, page 16-1.

SELinux on the Unified CM and Application Servers

Cisco Unified Communications System application servers that are based on the Unified CM platform use Security Enhanced Linux (SELinux) as the Host Intrusion Prevention software. For more information, see the section on Cisco Unified CM Security, page 4-21.

General Server Guidelines

Cisco Unified CM and other Collaboration application servers should not be treated as normal servers. Anything you do while configuring the system could affect calls that are trying to be places or that are in progress. As with any other business-class application, major configuration changes should be done within maintenance windows to keep from disrupting phone conversations.

Standard security policies for application servers might not be adequate for Collaboration servers. Unlike email servers or web servers, voice servers will not allow you to refresh a screen or re-send a message. The voice communications are real-time events. Any security policy for Collaboration servers should ensure that work that is not related to configuring or managing the voice systems is not done on the Collaboration servers at any time. Activities that might be considered normal on application servers within a network (for example, surfing the internet) should not take place on the Collaboration servers.

In addition, Cisco provides a well-defined patch system for the Collaboration servers, and it should be applied based on the patch policy within your IT organization. You should not patch the system normally using the OS vendor's patch system unless it is approved by Cisco Systems. All patches should be downloaded from Cisco or from the OS vendor as directed by Cisco Systems, and applied according to the patch installation process.

You should use the OS hardening techniques if your security policy requires you to lock down the OS even more than what is provided in the default installation.

To receive security alerts, you can subscribe to the Cisco Notification Service at:

https://www.cisco.com/go/support/

Deployment Examples

This section presents examples of what could be done from a security perspective for a lobby phone and a firewall deployment. A good security policy should be in place to cover deployments similar to these types.

Lobby Phone Example

The example in this section illustrates one possible way to configure a phone and a network for use in an area with low physical security, such as a lobby area. None of the features in this example are required for a lobby phone, but if your security policy states more security is needed, then you could use the features listed in this example.

Because you would not want anyone to gain access to the network from the PC port on the phone, you should disable the PC port on the back of the phone to limit network access (see PC Port on the Phone, page 4-26). You should also disable the settings page on the phone so that potential attackers cannot see the IP addresses of the network to which the lobby phone is connected (see Settings Access, page 4-28). The disadvantage of not being able to change the settings on the phone usually will not matter for a lobby phone.

Because there is very little chance that a lobby phone will be moved, you could use a static IP address for that phone. A static IP address would prevent an attacker from unplugging the phone and then plugging into that phone port to get a new IP address (see IP Addressing, page 4-4). Also, if the phone is unplugged, the port state will change and the phone will no longer be registered with Unified CM. You can track this event in just the lobby phone ports to see if someone is trying to attach to the network.

Using static port security for the phone and not allowing the MAC address to be learned would mean that an attacker would have to change his MAC address to that of the phone, if he were able to discover that address. Dynamic port security could be used with an unlimited timer to learn the MAC address (but never unlearn it), so that it would not have to be added. Then the switch port would not have to be changed to clear that MAC address unless the phone is changed. The MAC address is listed in a label on the bottom of the phone. If listing the MAC address is considered a security issue, the label can be removed and replaced with a "Lobby Phone" label to identify the device. (See Switch Port, page 4-6.)

A single VLAN could be used and Cisco Discovery Protocol (CDP) could be disabled on the port so that attackers would not be able to see any information from the Ethernet port about that port or switch to which it is attached. In this case, the phone would not have a CDP entry in the switch for E911 emergency calls, and each lobby phone would need either a label or an information message to local security when an emergency number is dialed.

A static entry in the DHCP Snooping binding table could be made because there would be no DHCP on the port (see DHCP Snooping: Prevent Rogue DHCP Server Attacks, page 4-8). Once the static entry is in the DHCP Snooping binding table, Dynamic ARP Inspection could be enabled on the VLAN to keep the attacker from getting other information about one of the Layer 2 neighbors on the network (see Requirement for Dynamic ARP Inspection, page 4-11).

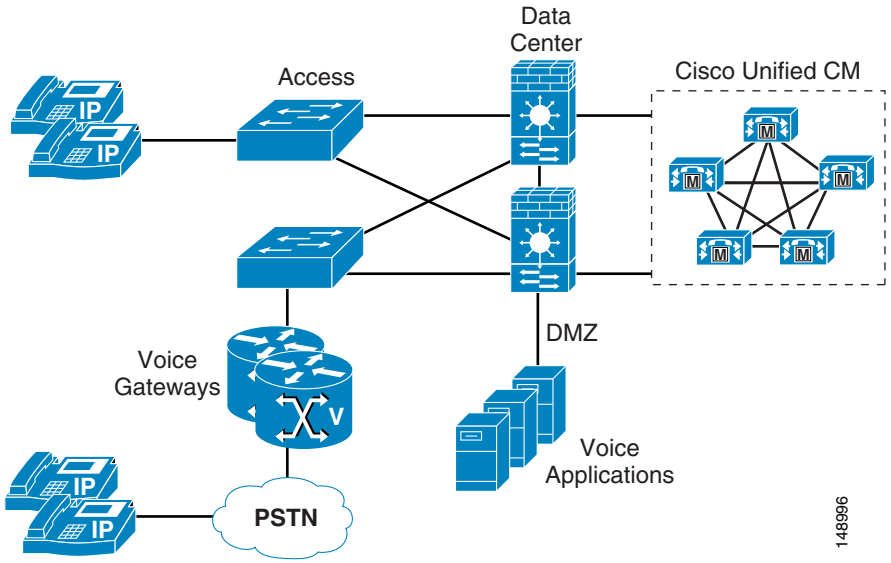
With a static entry in the DHCP Snooping binding table, IP Source Guard could be used. If an attacker got the MAC address and the IP address and then started sending packets, only packets with the correct IP address could be sent.

A VLAN ACL could be written to allow only the ports and IP addresses that are needed for the phones to operate (see VLAN Access Control Lists, page 4-32). The following example contains a very small ACL that can be applied to a port at Layer 2 or at the first Layer 3 device to help control access into the network (see Router Access Control Lists, page 4-32). This example is based on a Cisco 7960 IP Phone being used in a lobby area, without music on hold to the phone or HTTP access from the phone.

Firewall Deployment Example (Centralized Deployment)

The example in this section is one way that firewalls could be deployed within the data center, with Unified CMs behind them (see Figure 4-16). In this example, the Unified CMs are in a centralized deployment, single cluster with all the phones outside the firewalls. Because the network in this deployment already contained firewalls that are configured in routed mode within the corporate data center, the load was reviewed before the placement of gateways was determined. After reviewing the average load of the firewall, it was decided that all the RTP streams would not transverse the firewall in order to keep the firewalls under the 60% CPU load (see Putting Firewalls Around Gateways, page 4-39). The gateways are placed outside the firewalls, and ACLs within the network are used to control the TCP data flow to and from the gateways from the Unified CMs. An ACL is also written in the network to control the RTP streams from the phones because the IP addresses of the phones are well defined (see IP Addressing, page 4-4). The voice applications servers are placed within the demilitarized zone (DMZ), and ACLs are used at the firewalls to control access to and from the Unified CMs and to the users in the network. This configuration will limit the amount of RTP streams through the firewall using inspects, which will minimize the impact to the firewalls when the new voice applications are added to the existing network.

Figure 4-16 Firewall Deployment Example



Conclusion

This chapter did not cover all of the security that could be enabled to protect the voice and video data within your network. The techniques presented here are just a subset of all the tools that are available to network administrators to protect all the data within a network. On the other hand, even these tools do not have to be enabled within a network, depending on what level of security is required for the data within the network overall. Choose your security methods wisely. As the security within a network increases, so do the complexity and troubleshooting problems. It is up to each enterprise to define both the risks and the requirements of its organization and then to apply the appropriate security within the network and on the devices attached to that network.

Gateways

Revised: March 1, 2018

Gateways provide a number of methods for connecting a network of collaboration endpoints to the Public Switched Telephone Network (PSTN), a legacy PBX, or external systems. Voice and video gateways range from entry-level and standalone platforms to high-end, feature-rich integrated routers, chassis-based systems, and virtualized applications.

This chapter explains important factors to consider when selecting a Cisco gateway to provide the appropriate protocol and feature support for your voice and video network. The main topics discussed in this chapter include:

- Types of Cisco Gateways, page 5-2
- Cisco TDM and Serial Gateways, page 5-2
- Gateways for Video Telephony, page 5-11
- IP Gateways, page 5-15
- Best Practices for Gateways, page 5-32
- Fax and Modem Support, page 5-37

What's New in This Chapter

Table 5-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 5-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Minor updates for Cisco Expressway	Cisco Expressway, page 5-16	March 1, 2018

Types of Cisco Gateways

Until approximately 2006, the only way for an enterprise to connect its internal voice and video network to services outside the enterprise was by means of TDM or serial gateways to the traditional PSTN. Cisco offers a full range of TDM and serial gateways with analog and digital connections to the PSTN as well as to PBXs and external systems. TDM connectivity covers a wide variety of low-density analog (FXS and FXO), low density digital (BRI), and high-density digital (T1, E1, and T3) interface choices.

Starting around 2006, new voice and video service options to an enterprise became available from service providers, often as SIP trunk services. Using a SIP trunk for connecting to PSTN and other destinations outside the enterprise involves an IP-to-IP connection at the edge of the enterprise's network. The same functions traditionally fulfilled by a TDM or serial gateway are still needed at this interconnect point, including demarcation, call admission control, quality of service, troubleshooting boundary, security checks, and so forth. For voice and video SIP trunk connections, the Cisco Unified Border Element and the Cisco Expressway Series fulfill these functions as an interconnection point between the enterprise and the service provider network.

This chapter discusses in detail Cisco TDM and Serial gateway platforms and Cisco Expressway. Cisco Unified Border Element is also discussed briefly.

Cisco TDM and Serial Gateways

Cisco gateways enable voice and video endpoints to communicate with external telecommunications devices. There are two types of Cisco TDM gateways, analog and digital. Both types support voice calls, but only digital gateways support video.

Cisco Analog Gateways

There are two categories of Cisco analog gateways, station gateways and trunk gateways.

- Analog station gateways

  Analog station gateways connect Unified CM to Plain Old Telephone Service (POTS) analog telephones, interactive voice response (IVR) systems, fax machines, and voice mail systems. Station gateways provide Foreign Exchange Station (FXS) ports.

- Analog trunk gateways

  Analog trunk gateways connect Unified CM to PSTN central office (CO) or PBX trunks. Analog trunk gateways provide Foreign Exchange Office (FXO) ports for access to the PSTN, PBXs, or key systems, and E&M (recEive and transMit, or ear and mouth) ports for analog trunk connection to a legacy PBX. Analog Direct Inward Dialing (DID) and Centralized Automatic Message Accounting (CAMA) are also available for PSTN connectivity.

Cisco analog gateways are available on the following products and series:

- Cisco Analog Voice Gateways VG204XM and VG300 Series (VG310, VG320, VG350) all support SCCP.
- Cisco Integrated Services Routers Generation 2 (ISR G2) 2900, 3900, 3900E, and 4000 Series (4300 and 4400) with appropriate PVDMs and service modules or cards. PVDM4s utilized by ISR 4000 Series do not support video today.
- Cisco Analog Telephone Adapter (ATA) 190 (SIP only) provides a replacement for the ATA188.

Cisco Digital Trunk Gateways

A Cisco digital trunk gateway connects Unified CM to the PSTN or to a PBX via digital trunks such as Primary Rate Interface (PRI), Basic Rate Interface (BRI), serial interfaces (V.35, RS-449, and EIA-530), or T1 Channel Associated Signaling (CAS). Digital T1 PRI and BRI trunks can be used for both video and audio-only calls.

Cisco digital trunk gateways are available on the following products and series:

- Cisco Integrated Services Routers Generation 2 (ISR G2) 1900, 2900, 3900, 3900E, 4300, and 4400 Series with appropriate PVDMs and service modules or cards
- Cisco TelePresence ISDN GW 3241 and MSE 8321
- Cisco TelePresence Serial GW 3340 and MSE 8330

Cisco TelePresence ISDN Link

The Cisco TelePresence ISDN Link is a compact appliance for in-room ISDN and external network connectivity supporting Cisco TelePresence EX, MX, SX, and C Series endpoints. While traditional voice and video gateways are shared resources that provide connectivity between the IP network and the PSTN for many endpoints, each Cisco ISDN Link is paired with a single Cisco endpoint. For more information, refer to Cisco TelePresence ISDN Link documentation available at

https://www.cisco.com/en/US/products/ps12504/tsd_products_support_series_home.html

TDM Gateway Selection

When selecting a gateway for your voice and video network, consider the following factors:

- Gateway Protocols for Call Control, page 5-3
- Core Feature Requirements, page 5-5

Gateway Protocols for Call Control

Cisco Unified Communications Manager (Unified CM) supports the following IP protocols for gateways:

- Session Initiation Protocol (SIP)
- H.323
- Media Gateway Control Protocol (MGCP)
- Skinny Client Control Protocol (SCCP)

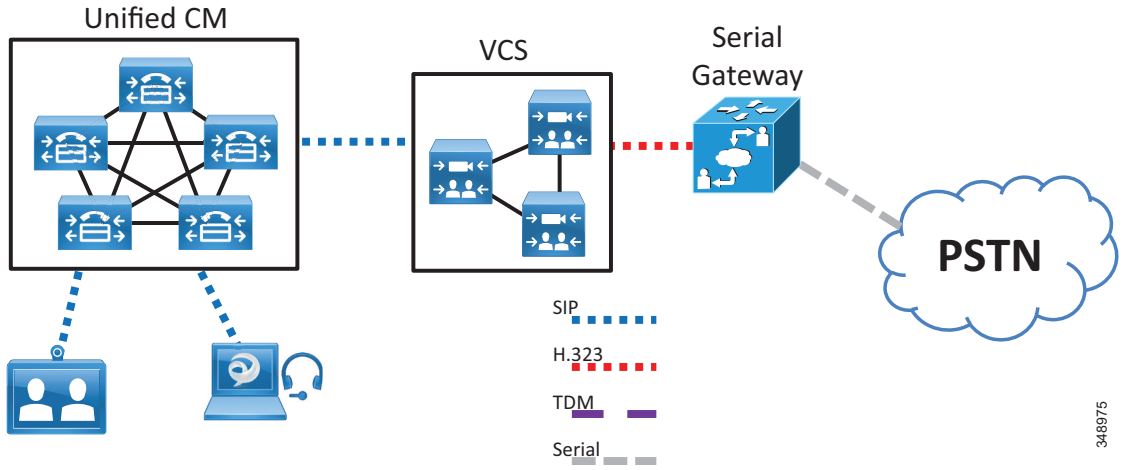
Cisco Expressway Series and Cisco TelePresence Video Communication Server (VCS) support the following IP protocols for gateways:

- Session Initiation Protocol (SIP)
- H.323

SIP is the recommended call signaling protocol because it aligns with the overall Cisco Collaboration solution and the direction of new voice and video products. However, protocol selection might depend on site-specific requirements and the current installed base of equipment. Existing deployments might be limited by the gateway hardware or require a different signaling protocol for a specific feature.

For example, placement of certain Cisco video gateways within the network depends upon the existing call control architecture. Both the Cisco ISDN and serial gateways are optimized for video calls and were initially designed to work with the Cisco VCS. The Cisco TelePresence Serial Gateway 8330 and 3340 platforms are recommended to register with a Cisco VCS using H.323, as shown in Figure 5-1.

Figure 5-1 Cisco TelePresence Serial Gateway Registered to Cisco VCS



The Cisco TelePresence ISDN Gateway 8321 and 3241 support SIP beginning with version 2.2. The Cisco 8321 and 3241 gateways can either register to VCS using H.323 (as shown in Figure 5-2) or trunk directly to Unified CM using SIP (as shown in Figure 5-3).

Figure 5-2 Cisco TelePresence ISDN Gateway Trunked to Cisco VCS

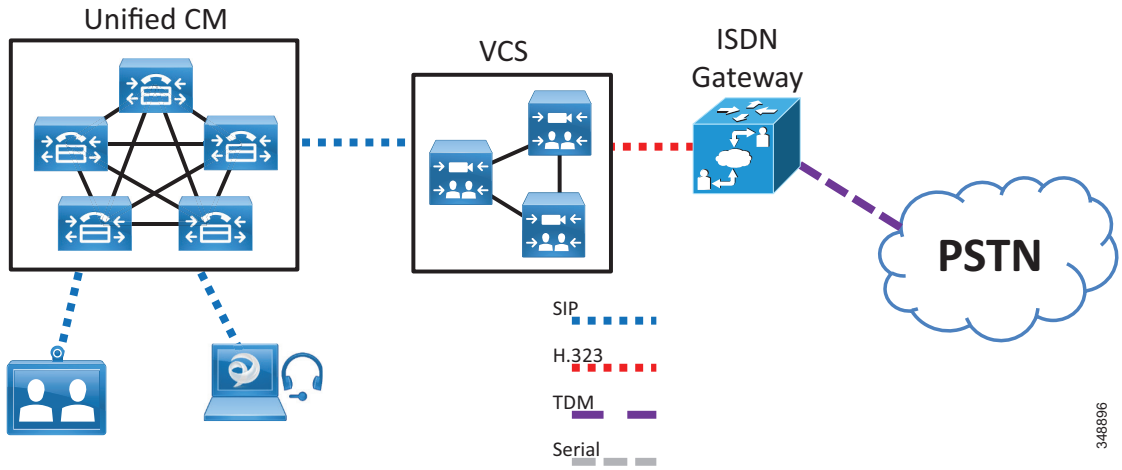
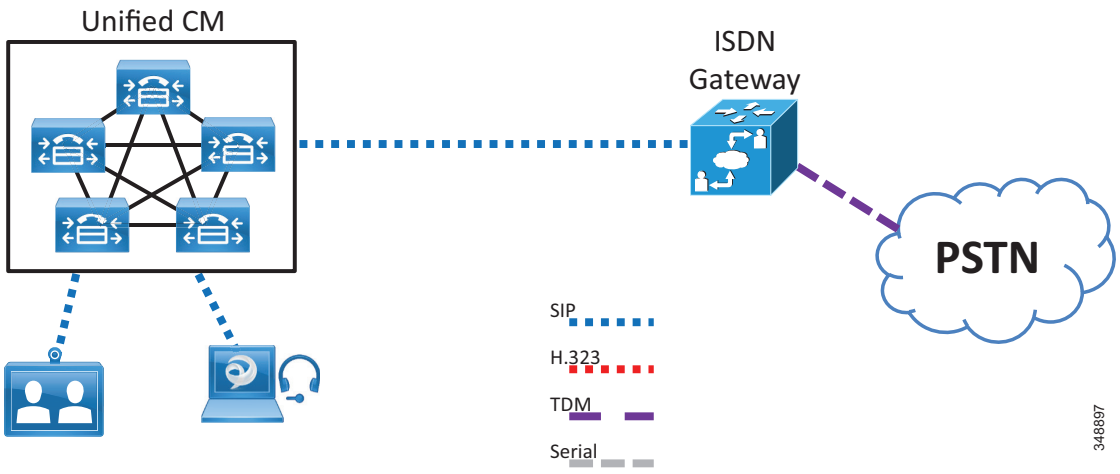


Figure 5-3 Cisco TelePresence ISDN Gateway Registered to Cisco Unified CM



In addition, the Unified CM deployment model being used can influence gateway protocol selection. (Refer to the chapter on Collaboration Deployment Models, page 10-1.)

Core Feature Requirements

Gateways used by voice and video endpoints must meet the following core feature requirements:

- DTMF Relay, page 5-5
- Supplementary Services, page 5-6

  Supplementary services are basic telephony functions such as hold, transfer, and conferencing.
- Unified CM Redundancy, page 5-9

  Cisco Unified Communications is based on a distributed model for high availability. Unified CM clusters provide for Unified CM redundancy. The gateways must support the ability to “re-home” to a secondary Unified CM in the event that a primary Unified CM fails. Some gateways may register to a Cisco VCS, in which case the gateway must support the ability to "re-home" to a secondary Cisco VCS if the primary fails.

Refer to the gateway product documentation to verify that any gateway you select for an enterprise deployment can support the preceding core requirements. Additionally, every collaboration implementation has its own site-specific feature requirements, such as analog or digital access, DID, and capacity requirements.

DTMF Relay

Dual-Tone Multifrequency (DTMF) is a signaling method that uses specific pairs of frequencies within the voice band for signals. A 64 kbps pulse code modulation (PCM) voice channel can carry these signals without difficulty. However, when using a low bite-rate codec for voice compression, the potential exists for DTMF signal loss or distortion. An out-of-band signaling method for carrying DTMF tones across an IP infrastructure provides an elegant solution for these codec-induced symptoms.

SCCP Gateways

The Cisco VG300 Series carries DTMF signals out-of-band using Transmission Control Protocol (TCP) port 2002. Out-of-band DTMF is the default gateway configuration mode for the VG310, VG320, and VG350.

H.323 Gateways

H.323 gateways, such as the Cisco 4000 Series products, can communicate with Unified CM using the enhanced H.245 capability for exchanging DTMF signals out-of-band. This capability is enabled through the command line interface (CLI) of the 4000 Series gateway and the **dtmf-relay** command available in its dial-peers.

MGCP Gateway

Cisco IOS-based platforms can use MGCP for Unified CM communication. Within the MGCP protocol is the concept of *packages*. The MGCP gateway loads the DTMF package upon start-up. The MGCP gateway sends *symbols* over the control channel to represent any DTMF tones it receives. Unified CM then interprets these signals and passes on the DTMF signals, out-of-band, to the signaling endpoint.

The method used for DTMF can be configured using the gateway CLI command:

mgcp dtmf-relay voip codec all mode {*DTMF method*}

Note An MGCP gateway cannot be forced to advertise only in-band DTMF. On enabling in-band DTMF relay, the MGCP gateway will advertise both in-band and out-of-band (OOB) DTMF methods. Unified CM determines which method should be selected and informs the gateway using MGCP signaling. If both the endpoints are MGCP, there is no ability to invoke in-band for DTMF relay because after enabling in-band DTMF, both sides will advertise in-band and OOB DTMF methods to Unified CM. Unified CM will always select OOB if in-band and OOB capabilities are supported by the endpoints.

SIP Gateway

Cisco IOS and ISDN gateways can use SIP for Unified CM communication. They support various methods for DTMF, but only the following methods can be used to communicate with Unified CM:

- Named Telephony Events (NTE), or RFC 2833
- Unsolicited SIP Notify (UN) (Cisco IOS gateways only)
- Key Press Markup Language (KPML)

The method used for DTMF can be configured in Cisco IOS using the gateway CLI command **dtmf-relay** under the respective **dial-peer**. The Cisco ISDN gateways support RFC 2833 and KPML for DTMF.

For more details on DTMF method selection, see the section on Calls over SIP Trunks, page 7-9.

Supplementary Services

Supplementary services provide user functions such as hold, transfer, and conferencing. These are considered basic telephony features and are more common in voice calls than in video calls.

SCCP Gateways

The Cisco SCCP gateways provide full supplementary service support. The SCCP gateways use the Gateway-to-Unified CM signaling channel and SCCP to exchange call control parameters.

H.323 Gateways

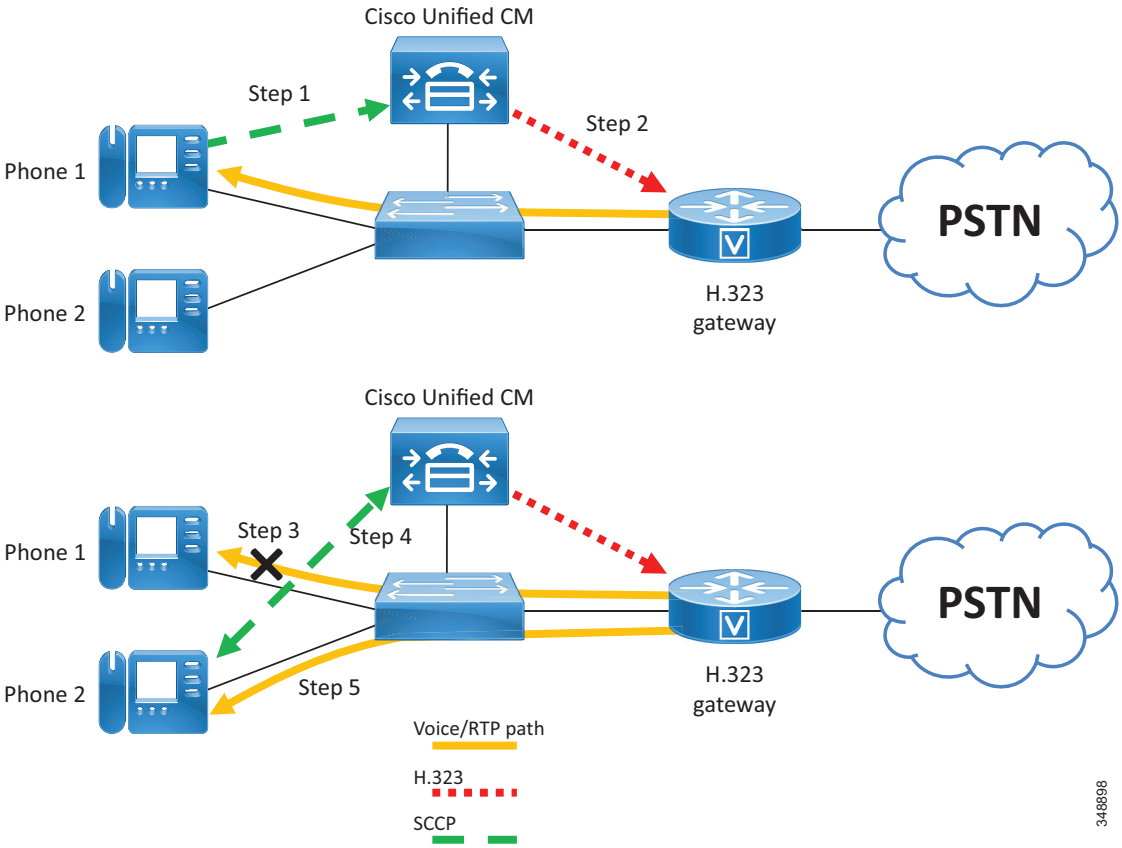
H.323v2 implements Open/Close LogicalChannel and the emptyCapabilitySet features. The use of H.323v2 by H.323 gateways eliminates the requirement for an MTP to provide supplementary services. A transcoder is allocated dynamically only if required during a call to provide access to G.711-only devices while still maintaining a G.729 stream across the WAN.

Once an H.323v2 call is set up between a Cisco IOS gateway and an IP endpoint, using the Unified CM as an H.323 proxy, the endpoint can request to modify the bearer connection. Because the Real-Time Transport Protocol (RTP) stream is directly connected to the endpoint from the Cisco IOS gateway, a supported media codec can be negotiated.

Figure 5-4 and the following steps illustrate a call transfer between two IP phones:

1. If IP Phone 1 wishes to transfer the call from the Cisco IOS gateway to Phone 2, it issues a transfer request to Unified CM using SCCP.
2. Unified CM translates this request into an H.323v2 CloseLogicalChannel request to the Cisco IOS gateway for the appropriate SessionID.
3. The Cisco IOS gateway closes the RTP channel to Phone 1.
4. Unified CM issues a request to Phone 2, using SCCP, to set up an RTP connection to the Cisco IOS gateway. At the same time, Unified CM issues an OpenLogicalChannel request to the Cisco IOS gateway with the new destination parameters, but using the same SessionID.
5. After the Cisco IOS gateway acknowledges the request, an RTP voice bearer channel is established between Phone 2 and the Cisco IOS gateway.

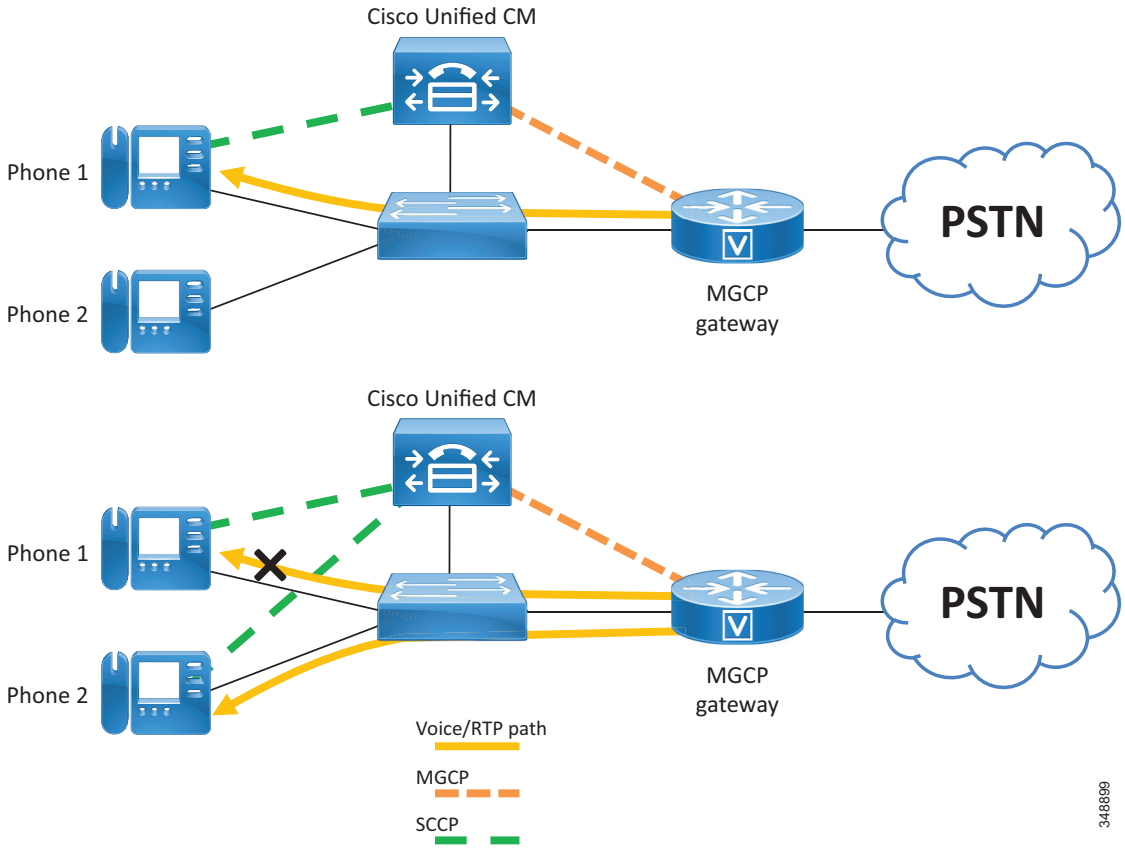
Figure 5-4 H.323 Gateway Supplementary Service Support



MGCP Gateway

The MGCP gateways provide full support for the hold, transfer, and conference features through the MGCP protocol. Because MGCP is a master/slave protocol with Unified CM controlling all session intelligence, Unified CM can easily manipulate MGCP gateway voice connections. If an IP telephony endpoint (for example, an IP phone) needs to modify the session (for example, transfer the call to another endpoint), the endpoint would notify Unified CM using SCCP. Unified CM then informs the MGCP gateway, using the MGCP User Datagram Protocol (UDP) control connection, to terminate the current RTP stream associated with the Session ID and to start a new media session with the new endpoint information. Figure 5-5 illustrates the protocols exchanged between the MGCP gateway, endpoints, and Unified CM.

Figure 5-5 MGCP Gateway Supplementary Service Support



SIP Gateway

The Unified CM SIP trunk interface to Cisco SIP gateways supports supplementary services such as hold, blind transfer, and attended transfer. The support for supplementary services is achieved via SIP methods such as INVITE and REFER. The corresponding SIP gateway must also support these methods in order for supplementary services to work. For more details, refer to the following documentation:

- *Cisco Unified Communications Manager System Guide*

  https://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html
- *Cisco IOS SIP Configuration Guide*

  https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/voice/sip/configuration/15-mt/sip-config-15-mt-book.html
- Cisco TelePresence ISDN Gateway documentation

  https://www.cisco.com/en/US/products/ps11448/tsd_products_support_series_home.html

Unified CM Redundancy

An integral piece of the collaboration solution architecture is the provisioning of low-cost, distributed PC-based systems to replace expensive and proprietary legacy PBX systems. This distributed design lends itself to the robust fault tolerant architecture of clustered Unified CMs. Even in its most simplistic form (a two-system cluster), a secondary Unified CM should be able to pick up control of all gateways initially managed by the primary Unified CM.

SCCP Gateways

Upon boot-up, the Cisco VG310, VG320, and VG350 gateways are provisioned with Unified CM server information. When these gateways initialize, a list of Unified CMs is downloaded to the gateways. This list is prioritized into a primary Unified CM and secondary Unified CM. In the event that the primary Unified CM becomes unreachable, the gateway registers with the secondary Unified CM.

H.323 VoIP Call Preservation for WAN Link Failures

H.323 call preservation enhancements for WAN link failures sustain connectivity for H.323 topologies where signaling is handled by an entity that is different from the other endpoint, such as a gatekeeper that provides routed signaling or a call agent (such as Cisco Unified CM) that brokers signaling between the two connected parties. Call preservation is useful when a gateway and the other endpoint are located at the same site but the call agent is remote and therefore more likely to experience connectivity failures.

H.323 call preservation covers the following types of failures and connections.

Failure Types:

- WAN failures that include WAN links flapping or degraded WAN links.
- Cisco Unified CM software failure, such as when the ccm.exe service crashes on a Unified CM server.
- LAN connectivity failure, except when a failure occurs at the local branch.

Connection Types:

- Calls between two Cisco Unified CM controlled endpoints under the following conditions:
  - During Unified CM reloads.
  - When a Transmission Control Protocol (TCP) connection used for signaling H.225.0 or H.245 messages between one or both endpoints and Unified CM is lost or flapping.
  - Between endpoints that are registered to different Unified CMs in a cluster, and the TCP connection between the two Unified CMs is lost.
  - Between IP phones and the PSTN at the same site.
- Calls between a Cisco IOS gateway and an endpoint controlled by a softswitch, where the signaling (H.225.0, H.245 or both) flows between the gateway and the softswitch and media flows between the gateway and the endpoint:
  - When the softswitch reloads.
  - When the H.225.0 or H.245 TCP connection between the gateway and the softswitch is lost, and the softswitch does not clear the call on the endpoint.
  - When the H.225.0 or H.245 TCP connection between softswitch and the endpoint is lost, and the softswitch does not clear the call on the gateway.
- Call flows involving a Cisco Unified Border Element running in media flow-around mode that reload or lose connection with the rest of the network.

Note that, after the media is preserved, the call is torn down later when either one of the parties hangs up or media inactivity is detected. In cases where there is a machine-generated media stream, such as music streaming from a media server, the media inactivity detection will not work and then the call might hang. Cisco Unified CM addresses such conditions by indicating to the gateway that such calls should not be preserved, but third-party devices or the Cisco Unified Border Element would not do this.

Flapping is defined for this feature as the repeated and temporary loss of IP connectivity, which can be caused by WAN or LAN failures. H.323 calls between a Cisco IOS gateway and Cisco Unified CM may be torn down when flapping occurs. When Unified CM detects that the TCP connection is lost, it clears the call and closes the TCP sockets used for the call by sending a TCP FIN, without sending an H.225.0 Release Complete or H.245 End Session message. This is called *quiet clearing*. The TCP FIN sent from Unified CM could reach the gateway if the network comes up for a short duration, and the gateway will tear down the call. Even if the TCP FIN does not reach the gateway, the TCP keepalives sent from the gateway could reach Unified CM when the network comes up. Unified CM will send TCP RST messages in response to the keepalives because it has already closed the TCP connection. The gateway will tear down H.323 calls if it receives the RST message.

Configuration of H.323 call preservation enhancements for WAN link failures involves configuring the **call preserve** command. If you are using Cisco Unified CM, you must enable the Allow Peer to Preserve H.323 Calls parameter from the Service Parameters window.

The **call preserve** command causes the gateway to ignore socket closure or socket errors on H.225.0 or H.245 connections for active calls, thus allowing the socket to be closed without tearing down calls using those connections.

MGCP Gateway

MGCP gateways also have the ability to fail over to a secondary Unified CM in the event of communication loss with the primary Unified CM. When the failover occurs, active calls are preserved.

Within the MGCP gateway configuration file, the primary Unified CM is identified using the **call-agent <hostname>** command, and a list of secondary Unified CM is added using the **ccm-manager redundant-host** command. Keepalives with the primary Unified CM are through the MGCP

application-level keepalive mechanism, whereby the MGCP gateway sends an empty MGCP notify (NTFY) message to Unified CM and waits for an acknowledgement. Keepalive with the backup Unified CMs is through the TCP keepalive mechanism.

If the primary Unified CM becomes available at a later time, the MGCP gateway can "re-home," or switch back to the original Unified CM. This re-homing can occur either immediately, after a configurable amount of time, or only when all connected sessions have been released.

SIP Gateway

Redundancy with Cisco IOS SIP gateways can be achieved similarly to H.323. If the SIP gateway cannot establish a connection to the primary Unified CM, it tries a second Unified CM defined under another dial-peer statement with a higher preference.

By default the Cisco IOS SIP gateway transmits the SIP INVITE request 6 times to the Unified CM IP address configured under the dial-peer. If the SIP gateway does not receive a response from that Unified CM, it will try to contact the Unified CM configured under the other dial-peer with a higher preference.

Cisco IOS SIP gateways wait for the SIP 100 response to an INVITE for a period of 500 ms. By default, it can take up to 3 seconds for the Cisco IOS SIP gateway to reach the backup Unified CM. You can change the SIP INVITE retry attempts under the **sip-ua** configuration by using the command **retry invite** <*number*>. You can also change the period that the Cisco IOS SIP gateway waits for a SIP 100 response to a SIP INVITE request by using the command **timers trying** <*time*> under the **sip-ua** configuration.

One other way to speed up the failover to the backup Unified CM is to configure the command **monitor probe icmp-ping** under the **dial-peer** statement. If Unified CM does not respond to an Internet Control Message Protocol (ICMP) echo message (ping), the dial-peer will be shut down. This command is useful only when the Unified CM is not reachable. ICMP echo messages are sent every 10 seconds.

The Cisco ISDN Gateway can connect to Unified CM via SIP trunk starting with Unified CM release 9.0 and ISDN Gateway release 2.2 and later. The ISDN Gateway SIP configuration consists of entering an IP address, hostname, DNS A record, or DNS SRV record for outbound SIP connections. Redundancy can be achieved by utilizing DNS SRV records with appropriate weight and priority so that, if the primary Unified CM fails, the ISDN Gateway will send outbound SIP calls to the secondary Unified CM.

Gateways for Video Telephony

Video gateways terminate video calls into an IP telephony network or the PSTN. Video gateways are different from voice gateways because they have to interact with the ISDN or serial links that support video and convert that call to a video call on the IP network using protocols such as H.323 or SIP. Enterprises can consider separate gateways for voice calls and video calls, or they can have integrated gateways that route both voice and video calls.

The following key considerations can help you decide if you need separate gateways for voice and video or an integrated gateway:

- Dial plan — If the enterprise has the flexibility of a separate dial plan for video users, it can use separate video gateways that allow it to keep existing enterprise dial plans.
- Video users — If the enterprise has a large number of users who primarily use voice rather than video, then Cisco recommends using separate video gateways to service the video call users.
- Locations — If the enterprise has a large number of distributed locations with video users at many locations, then Cisco recommends using an integrated gateway to reduce total cost of ownership (TCO).

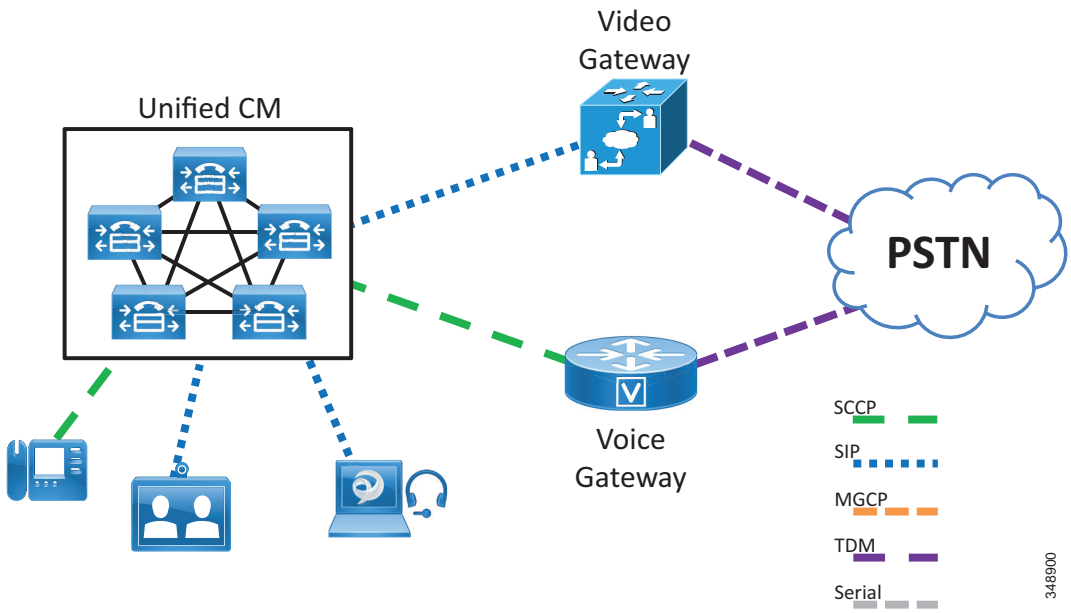
- Additional video capabilities such as video IVR, auto attendant, and bonding across trunks — Dedicated video gateways support advanced features that integrated gateways do not support.
- Protocol — Gateway protocol can be an important factor to align with enterprise policies and standards.
- Device management — Ease of maintenance, management, and troubleshooting can be an important factor. Dedicated gateways provide a better user interface (GUI) for management and configuration, while integrated gateways can provide better troubleshooting. However, these factors are dependent on the respective products.

Dedicated Video Gateways

Enterprises that have an extensive voice infrastructure with voice gateways can add dedicated video gateways so that users can make video calls through them to the PSTN. The Cisco ISDN Gateway and Serial Gateways are examples of dedicated video gateways. Although these products support audio-only calls, they were designed specifically with video users in mind. They support a wide range of video-centric protocols and features.

Figure 5-6 shows an enterprise deployment that can use existing protocols for its voice gateways and add video gateways so that Unified CM users can make voice and video calls to the PSTN.

Figure 5-6 Unified CM System with Separate PSTN Lines for Voice and IP Video Telephony



The Cisco video gateways, while excellent for video calls, do not support all of the telephony features that Cisco voice gateways offer. Cisco video gateways have the following characteristics:

- The Serial Gateway supports only H.323 for IP connectivity.
- The ISDN Gateway supports H.323 and SIP (starting with release 2.2) for IP connectivity.
- They support T1/E1-PRI, BRI, V.35, RS-449, and EIA-530.
- They support H.261, H.263, H.263+, and H.264 video codecs.

- They support G.711, G.722, G.722.1, and G.728; but they do not support G.729 audio.
- They support H.320, H.233, H.234, H.235 (AES), H.239, H.221, FTP, RTP, HTTP, HTTPS, DHCP, SNMP, and NTP.

As a result of these differences in the products, the Cisco TDM and Serial Gateways are not recommended as replacements for Cisco voice gateways. IP Telephony customers who want to add video to their communications environment should deploy both types of gateways and use the Cisco voice gateways for all voice calls and use the Cisco video gateways for video calls only. Customers might also have to procure separate circuits for voice and video from their PSTN service provider, depending on which model of Cisco gateway is deployed.

Also consider how calls will be routed across the IP network to a remote gateway for the purpose of providing toll bypass, and how calls will be re-routed over the PSTN in the event that the IP network is unavailable or does not have enough bandwidth to complete the call. More specifically, do you want to invoke automated alternate routing (AAR) for video calls?

Integrated Video Gateways

Although not recommended, enterprises may consider an integrated device for voice and video gateway functionality. This provides the enterprise the advantages of managing fewer devices and keeping the dial plan simple. The gateway processes the call as a voice call if it is voice and as a video call if it is video.

Cisco IOS, ISDN, and Serial Video gateway have the following characteristics:

- Provide H.323 and SIP support (except Serial Gateway, which is H.323 only)
- Supports H.261, H.263, H.263+, and H.264 video codec
- Provides extensive called and calling transformation capabilities
- Provides extensive logging and troubleshooting capabilities

The following considerations apply for deploying Cisco IOS, ISDN, and Serial Video gateways:

- Consider the capacity needed on PSTN links for additional video calls.
- Consider the need of devices to use content sharing such as Binary Floor Control Protocol (BFCP), and the additional bandwidth that will be used on the IP network.
- Consider if users need features such as far-end camera control or DTMF that is used for conferences that the gateway needs to support.

Configuring Video Gateways in Unified CM

You can configure a Cisco TelePresence ISDN Gateway in either of the following ways:

- Configure a SIP trunk pointing to the ISDN gateway (as shown in Figure 5-3), and add appropriate Unified CM route patterns pointing to the SIP trunk.
- Configure a SIP trunk from Unified CM to Cisco VCS. Have the ISDN gateway (or Serial gateway in this case) register to the VCS using H.323 (as shown in Figure 5-2.

The Cisco TelePresence Serial Gateway cannot be trunked directly to Unified CM. It must register to Cisco VCS, which in turn has a SIP trunk to Unified CM.

Either way, the goal is have all inbound calls received by the gateways sent to Unified CM so that Unified CM can decide how to route the calls. See the chapter on Cisco Unified CM Trunks, page 6-1, for more details on how to configure the SIP trunk between Unified CM and VCS.

Call Signaling Timers

Due to the delay inherent in H.320 bonding, video calls can take longer to complete than voice calls. Several timers in Unified CM are tuned, by default, to make voice calls process as fast as possible, and they can cause video calls to fail. Therefore, you must modify the following timers from their default values in order to support H.320 gateway calls:

- H.245TCSTimeout
- Media Exchange Interface Capability Timer
- Media Exchange Timer

Cisco recommends that you increase each of these timers to 25 by modifying them under the Service Parameters in Unified CM Administration. Note that these are cluster-wide service parameters, so they will affect calls to all types of devices, including voice calls to existing Cisco voice gateways.

Bearer Capabilities of Cisco IOS Voice Gateways

H.323 calls use the H.225/Q.931 Bearer Capabilities Information Element (bearer-caps) to indicate what type of call is being made. A voice-only call has its bearer-caps set to **speech** or **3.1 KHz Audio**, while a video call has its bearer-caps set to **Unrestricted Digital Information**. Some devices do not support Unrestricted Digital Information bearer-caps. Calls to these devices might fail if Unified CM attempts the call as a H.323 video call.

Unified CM decides which bearer-caps to set, based on the following factors:

- Whether the calling and/or called devices are video-capable
- Whether the region in Unified CM is configured to allow video for calls between those devices

Unified CM supports retrying the video call as audio, and this feature can be enabled through configuration. When Unified CM makes a video call with bearer-caps set to **Unrestricted Digital** and the call fails, Unified CM then retries the same call as an audio call with the bearer-caps set to **speech**.

When using H.323, Cisco IOS gateways can service calls as voice or video, based on the bearer capabilities it receives in the call setup. When using SIP, the gateway translates the ISDN capabilities into SDP for call negotiations.

If the Cisco voice gateway uses MGCP to communicate with Unified CM, the problem will not occur because Unified CM does not support video on its MGCP protocol stack and because, in MGCP mode, Unified CM has complete control over the D-Channel signaling to the PSTN.

IP Gateways

The Cisco IP gateways Include:

- Cisco Unified Border Element, page 5-15
- Cisco Expressway, page 5-16

Cisco Unified Border Element

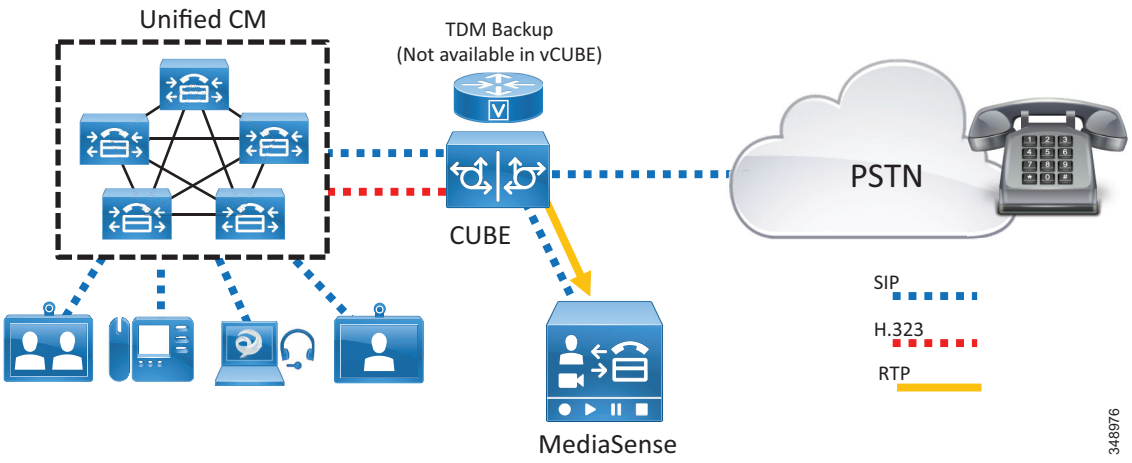
Innovations in collaboration services have delivered significant improvements in employee productivity, and enterprises are widely deploying IP-based Unified Communications, for both internal calling within the enterprise and external PSTN access. This has resulted in significant migration from TDM-based circuits, by both enterprises and telephony service providers, to IP-based trunks for Unified Communications. At the heart of IP-based telephony trunks lies the Session Initiation Protocol (SIP), which is an industry standard communications protocol based on RFC 3261 and is widely used for controlling multimedia communication sessions and applications such as voice, video, unified messaging, voicemail, and conferencing.

These PSTN SIP trunks terminate on a session Border Controller (SBC) at the enterprise, which serves as a demarcation point between the enterprise and the service provider IP networks, similar to how firewalls separate two data networks. The Cisco Unified Border Element (CUBE) Enterprise is Cisco's SBC offering, and it enables rich multimedia communications for enterprises by providing:

- Session Control — Call admission control, trunk routing, QoS, statistics, billing, redundancy, scalability, voice quality monitoring
- Security — Encryption, authentication, registration, SIP protection, voice policy, toll fraud prevention, telephony denial of service (TDoS) attack protection
- Interworking — Various SIP and H323 stack interoperability, sip normalization, dtmf, Transcoding, Transrating, Codec Filtering
- Demarcation — Fault isolation, topology and address hiding, L5/L7 protocol demarcation, network border

CUBE provides essential capabilities that ensure interoperability, security, and service assurance when carrying IP traffic via SIP trunks across various enterprises and service provider networks. It is a Back-to-Back User Agent (B2BUA) and is part of the Cisco IOS infrastructure on Cisco ISR G2 800 Series platforms, Cisco IOS-XE for the ASR 1000 Series, Cisco ISR 4000 Series, and CUBE on the Cisco Cloud Services Router (CSR) 1000V Series (virtual CUBE, or vCUBE). Figure 5-7 illustrates the enterprise CUBE deployment.

Figure 5-7 Cisco Unified Border Element Deployment



For more information about Cisco Unified Border Element, refer to the documentation at

https://www.cisco.com/go/cube

Cisco Expressway

Use of the Internet for collaboration services continues to increase in popularity and is quickly replacing existing legacy ISDN video systems and gateways. The two primary protocols leveraged for Internet based collaboration services are SIP and H.323. The Internet is also used to connect remote and mobile users to voice, video, IM and presence, and content sharing services without the use of a virtual private network (VPN).

The Expressway-C and Expressway-E pair performs the following functions:

- Mobile and remote access, as well as business-to-business services, can be enabled as part of the same Cisco Expressway-C and Expressway-E solution pair.
- Interworking — The capability to interconnect H.323-to-SIP calls for voice, video, and content sharing.
- Boundary communication services — While Expressway-C sits in the corporate network, Expressway-E is in the enterprise DMZ and provides a distinct connection point for communication services between the enterprise network and the Internet.
- Security — The capability to provide authentication and encryption for both mobile and remote access and business-to-business communications.

Expressway-C and Expressway-E are designed to work together to form a firewall traversal solution that is the core component for business-to-business communications over the Internet. Expressway-C sits on the inside (trusted side) of the enterprise network and serves the role of providing a secure, trusted, and standards-based way of connecting to Expressway-E. It acts as a traversal client to all devices behind it. This solves the problem for devices using a large number of media ports by multiplexing all media to a very small number of ports opened for outbound communications. It provides an authenticated and trusted connection from inside the enterprise to outside by sending a keep-alive for the traversal zone from Expressway-C to Expressway-E. Additionally, it provides a single point of contact for all Internet communications, thus minimizing the security risk.

Real-time and near real-time communication protocols such as SIP, H.323, and XMPP do not address the need to communicate with devices that might be behind a firewall. Typical communications using these protocols include the device IP address in the signaling and media, which becomes the payload of the TCP and UDP packets, respectively. When these devices are on the same internally routable network, they can successfully communicate directly with each other. The signaling IP address carried in the payload of the TCP packet is routable back to the initiating device, and vice versa. However, when the initiating device is on a different network behind a public or network edge firewall, two problems are encountered. The first problem is that the receiving device, after decoding the packet, will respond to the internal IP address carried in the payload. This IP address is typically a non-routable RFC 1918 address and will never reach the return destination. The second problem is that, even if the return IP address is routable, the media (which is RTP/UDP) is blocked by the external firewall. This applies to both business-to-business and mobile and remote access communications.

Expressway-E sits at the network edge in the DMZ. It serves the role of solving both the signaling and media routing problems for SIP, H323, and XMPP, while maintaining standards interoperability. Expressway-E changes the appropriate headers and IP addresses to process the media and signaling on behalf of the endpoints, devices, and application servers that are inside the network.

Expressway-C and Expressway-E Deployment for Business-to-Business Communications

The standard deployment of the Cisco Expressway Series involves deploying at least one Expressway-C and Expressway-E pair for business-to-business communications. Both Expressway-C and Expressway-E should be deployed in a cluster to provide better resiliency. The number of servers for each cluster depends on the number of concurrent calls. (For details, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.)

Frequently, multiple pairs of Expressway-C and Expressway-E are deployed for geographic coverage and scale, providing access to multiple instances of collaboration services. Unified CM is connected to Expressway-C through a SIP trunk for unified business communications access over the Internet. Based on the enterprise security policy, a number of different deployment models can be implemented. In this document we focus on a DMZ deployment of Expressway-E with dual network interfaces because it is the most common and secure deployment model. For additional deployment models, refer to latest version of the *Cisco Expressway Basic Configuration Deployment Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html

Expressway-C and Expressway-E provide firewall traversal capabilities. Firewall traversal works as follows:

1. Expressway-E is the traversal server installed within the enterprise DMZ, and Expressway-C is the traversal client installed inside the enterprise network.
2. Expressway-C initiates traversal connections outbound through the firewall to specific ports on Expressway-E, with secure login credentials. If the firewall allows outbound connections, as it does in the vast majority of cases, no additional ports are required to be opened in the enterprise firewall.

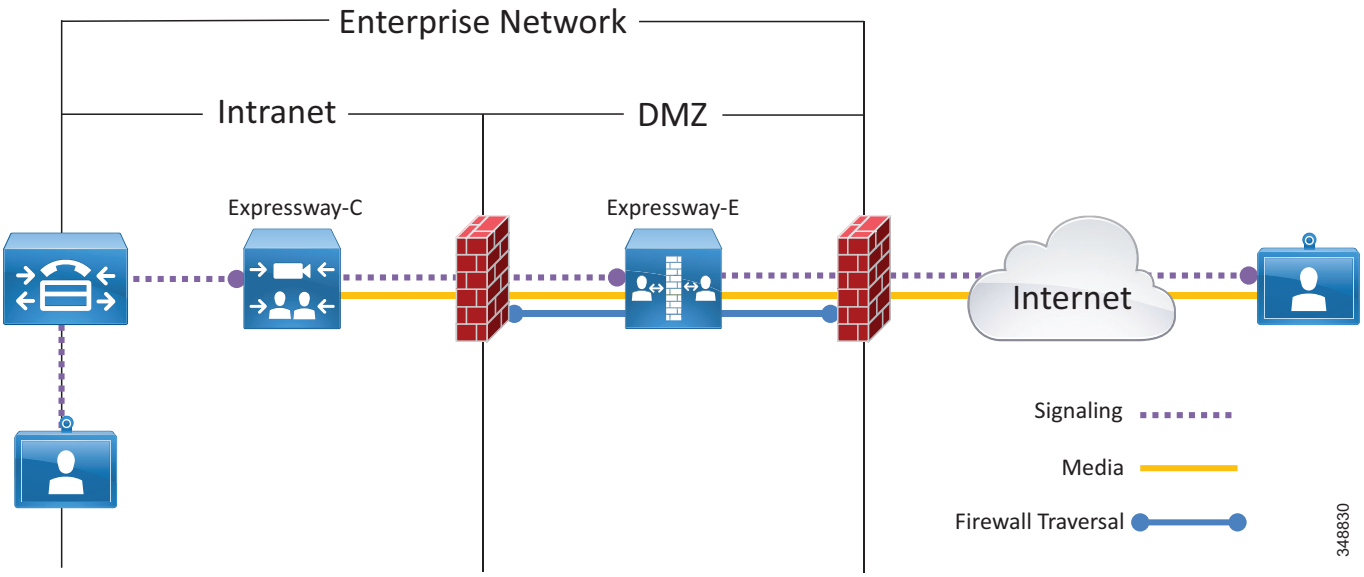
   For port details, refer to the latest version of the *Unified Communications Mobile and Remote Access via Cisco Expressway Deployment Guide*, which includes all ports used by Expressway in business-to-business and mobile and remote access scenarios. This guide is available at

   https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html
3. Once the connection has been established, Expressway-C sends periodic keep-alive packets to Expressway-E to maintain the connection.

4. When Expressway-E receives an incoming call or other collaboration service request, it issues an incoming request to Expressway-C.
5. Expressway-C then routes the request to Unified CM or other collaboration service applications.
6. The connection is established, and application traffic (including voice and video media) traverses the firewall securely over an existing traversal connection.

For firewall traversal to work, a traversal client zone has to be configured on Expressway-C and a traversal server zone has to be configured on Expressway-E. Figure 5-8 summarizes the firewall traversal process in a dual-interface deployment scenario for Expressway-E.

Figure 5-8 Firewall Traversal in a Dual-Interface Deployment



In the dual-interface deployment scenario, Expressway-E sits in the DMZ between two firewalls: the Internet firewall provides for NAT services toward the Internet, and the intranet firewall provides access to the corporate trusted network.

Expressway-E has two LAN interfaces: one toward the Internet firewall (also called the external interface) and the other toward the intranet firewall (also called the internal interface). In order to route packets to the external or internal interface, you create static routes on Expressway-E. The easiest way to create the static routes is by setting the Expressway-E default gateway equal to the default gateway for the external LAN interface, and by creating static routes for every internal network. In this way, internal traffic will be sent to the internal interface, and all traffic not matching the network range configured in the static routes will be sent to the Internet.

There is no need for the external interface to be assigned a public IP address because the address can be translated statically by NAT. In this case, the public IP address has to be configured on Expressway-E itself. The Expressway-E external interface can be statically translated by NAT, but the Expressway-E internal interface can be statically translated by NAT only if the Expressway is not clustered. The Expressway-C interface can be translated by NAT.

A connection from the Internet for business-to-business communications between Expressway-C and back-end application services may or may not be encrypted, based on the configuration and dictated by the corporate policies. Note that in this case the communication will be encrypted end-to-end only if both

the corporate and the remote business-to-business party supports encryption with public certificates. In all other cases, the video call will be sent unencrypted, or it will be dropped based on Expressway-E configuration policies.

Business-to-Business Call Flow

Business-to-business communications require the ability to look up the domains of remote organizations for the purpose of URI routing. This is done by creating a DNS zone on Expressway-E. This zone should be configured with the default settings. Both SIP and H.323 are set by default. Expressway-C and Expressway-E use the protocol that was used to initiate the call, and they automatically try the other protocol when SIP-to-H.323 gateway interworking is enabled on Expressway.

SIP-to-H.323 interworking should be set to **On** for Expressway-E. If a call is receive as an H.323 call, this allows Expressway-E to interwork the call to SIP and use native SIP for the rest of the call legs to Unified CM. Likewise, an outbound call to an H.323 system will remain a SIP call until it reaches Expressway-E, where it will be interworked to H.323.

In order to receive business-to-business communications over the Internet, External SIP and H.323 DNS records are required. These records allow other organizations to resolve the domain of the URI to the Expressway-E that is offering that call service. Cisco's validated design includes the SIP and SIPS SRV records and the H.323 SRV record for business-to-business communications. The H.323 SRV record is not necessary for Expressway-E because this record is used by an endpoint to find its gatekeeper for registration.

Table 5-2 shows the DNS SRV records used for resolving the domain of the URI.

Table 5-2 DNS SRV Records for Resolving the URI Domain

Type of Communication	Domain	Port	Protocol
SIP business-to-business	_sips._tcp.domain	5061	TLS
	_sip._tcp.domain	5060	TCP
	_sip._udp.domain	5060	UDP
H.323 business-to-business	_h323ls._udp.domain	1719	RAS
	_h323cs._tcp.domain	1720	H.225
Mobile and remote access	_collab-edge._tls.domain	8443	Jabber login
	_xmpp-server._tcp.domain	5269	XMPP Federation

For more information about configuring a DNS zone on Expressway-E, refer to latest version of the *Cisco Expressway Basic Configuration Deployment Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html

Outbound calls use a SIP Route Pattern on Cisco Unified CM set to "*". Any SIP URI that does not find a match inside the local Unified CM cluster or ILS table will be sent through this SIP Route Pattern, according to the routing rules logic defined in the chapter on Dial Plan, page 14-1. Configure this SIP Route Pattern to have a Route List to the Expressway-C cluster as a target.

Configure Expressway-C to have two rules for business-to-business communications:

- Send any SIP URI with the domain portion matching the domain of the company to Cisco Unified Communications Manager.
- Send any SIP URI with the domain portion matching any other domain to Expressway-E.

On Expressway-E configure two rules for business-to-business communications:

- Send any SIP URI with the domain portion matching the domain of the company to the Expressway-C cluster.
- Send any SIP URI with the domain portion matching any other domain to the DNS Zone that is used for DNS SRV resolution.

When a user dials a string followed by an external domain from an endpoint connected to Unified CM, the SIP Route Pattern will be matched. Unified CM will send the call to Expressway-C, and Expressway-C to will send it Expressway-E. Expressway-E will perform a DNS SRV lookup on a public DNS. The DNS will resolve the SRV record, and Expressway-E will be able to direct the call to the unknown remote edge.

Inbound calls will be received by the Expressway-E on the Default Zone, and based on the search rules specified above, Expressway-E will send the call to the Expressway-C, which will send it to Cisco Unified CM.

Note that any Cisco endpoint connected to Cisco Unified CM, regardless from model type or voice/video capabilities, will be reachable.

If the endpoint does not have any associated SIP URI, it will be reachable through the string *<DN>*@*<domain>*, where *<DN>* is the directory number configured on Cisco Unified CM and *<domain>* is the company SIP domain.

In case the device has a corresponding alphanumeric SIP URI associated with its DN, the same device can also be reached by dialing the alphanumeric SIP URI.

IP-Based Dialing for Business-to-Business Calls

IP-based dialing is a feature well known and used in most scenarios, especially when dealing with H.323 endpoints. The Cisco Collaboration Architecture uses SIP URIs and does not need IP-based dialing. However, when interacting with endpoints in other organizations that are capable of making and receiving calls using IP addresses only, the Cisco Collaboration Architecture allows IP-based dialing for both inbound and outbound calls.

Outbound Calls

Outbound IP dialing is supported on Expressway-E and Expressway-C, but it does not have full native support on Cisco Unified CM. However, it is possible to set up Unified CM to have IP-based dialing, as described here.

Instead of dialing the IP address alone, users on Cisco Unified CM can dial a SIP URI-based IP address as shown in this example: 10.10.10.10@ip, where "@ip" is literal and could be replaced with "external", "offsite" or other meaningful terms.

Unified CM will match a SIP route pattern configured to route the "ip" fictional domain to Expressway-C. Expressway-C strips off the domain "@ip" and sends the call to Expressway-E, which is also configured for IP address dialing.

Calls to unknown IP addresses on Expressway -E should be set to **Direct**. Since IP-based address dialing is mostly configured in H.323 endpoints when no call control is deployed, this allows Expressway-E to send H.323 calls directly to an endpoint at a public IP address. The call will remain a SIP call until interworked on Expressway-E.

Alternatively, instead of having to append the fictional domain, users might replace the dots with a star character, as in this example: 10*10*10*10.

Unified CM will match a Route Pattern defined as !*!*!*! and send the call to Expressway-C, which will replace the "star" character with a dot. In this case, the search rule will match the following regex expression: (\d\d?\d?)(*)(\d\d?\d?)(*)(\d\d?\d?)(*)(\d\d?\d?), and will have \1\.\3\.\5\.\7 as the replacement string.

Inbound Calls

IP-based inbound calls make use of a fallback alias configured in Expressway-E. When a user on the Internet dials the IP address of the Expressway-E external LAN interface, Expressway-E receives the call and sends the call to the alias configured in the fallback alias setting. As an example, if the fallback alias is configured to send the call to conference number 80044123 or to the conference alias meet@example.com, the inbound call will be sent to the application or device in charge of such conferences.

If the static mapping between the IP address and the fallback alias is too limited, it is possible to set the fallback alias to the pilot number of Cisco Unity Connection or Cisco Unified Contact Center Express (UCCX). In this way it is possible to use the Unity Connection auto-attendant or UCCX IVR feature to specify the final destination through DTMF, or by speech recognition if Unity Connection is enabled to support this feature. If Unity Connection is used as an auto-attendant feature for external endpoints dialing the IP address of the Expressway-E, remember to set the Rerouting Calling Search Space on the Unified CM trunk configuration for Unity Connection.

High Availability for Expressway-C and Expressway-E

We recommend deploying Expressway-C and Expressway-E in clusters. Each cluster can have up to six Expressway nodes and a maximum of N+2 physical redundancy. All nodes are active in the cluster. For details about cluster configuration, refer to the latest version of the *Cisco Expressway Cluster Creation and Maintenance Deployment Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html

Expressway clusters provide configuration redundancy. The first node configured in the cluster is the master. Configuration is done in the master and automatically replicated to the other nodes. Expressway clusters provide call license sharing and resilience. All rich-media session licenses are shared equally across nodes in the cluster. Call licenses are contributed by the licenses configured on each node.

Expressway-C and Expressway-E deployed as virtual machines support VMware VMotion. VMware. VMotion enables the live migration of running virtual machines from one physical server to another. When moving a virtual machine, Expressway-C and Expressway-E servers will maintain active calls when handling signaling only or when handling both signaling and media. This provides high availability for the Expressway nodes as well as call resilience across Cisco Unified Computing System (UCS) hosts.

The following rules apply to Expressway clustering:

- Expressway-C and Expressway-E node types cannot be mixed in the same cluster.
- All nodes in a cluster must have identical configurations.
- Configuration changes should be made only on the master node, and this will overwrite the configuration on the other nodes in the cluster when replication occurs.
- If a node becomes unavailable, the licenses it contributed to the cluster will become unavailable after 2 weeks.
- Deploy an equal number of nodes in Expressway-C and Expressway-E clusters.
- Deploy the same OVA template throughout the cluster.

- All nodes in a cluster need to be within 30 ms maximum round-trip time to all other cluster nodes. Clustering over the WAN is therefore not recommended due to latency constraints.
- You must use the same cluster preshared key for all nodes within the same cluster.
- If mobile and remote access and business-to-business communications are enabled on the same Expressway-C and Expressway-E pairs, the SIP port number used on the SIP trunk between Unified CM and Expressway-C must be changed from the default 5060 or 5061 (for example, use 5560 and 5561).
- A DNS SRV record must be available for the cluster and must contain A or AAAA records for each node of the cluster.

Since Expressway-C is deployed in the internal network and Expressway-E is in the DMZ, Expressway-C has to be connected to Expressway-E through a traversal zone for business-to-business calls. Mobile and remote access requires a separate traversal zone, referred to as the **Unified Communication traversal zone**. The traversal server and traversal client zones include all the nodes of Expressway-C and Expressway-E, so that if one of the nodes is not reachable, another node of the cluster will be reached instead.

The traversal client zone configured on Expressway-C should contain the fully qualified domain names of all of the cluster nodes of the corresponding Expressway-E cluster. Likewise, the traversal server zone should connect to all Expressway-C cluster nodes. This is achieved by including, in the subject alternative names of the Expressway-C certificate, the FQDNs of the Expressway-C cluster nodes and by setting the TLS verify subject name equal to the FQDN of the Expressway-C cluster. This creates a mesh configuration of cluster nodes across the traversal zone and provides continuous and high availability of the traversal zone until the last cluster node is unavailable.

Expressway-C connects to Unified CM via a neighbor zone for routing inbound and outbound business-to-business calls. Unified CM also trunks to Expressway-C. For high availability, the fully qualified domain names of each Expressway-C cluster node should be listed in the trunk configuration on Unified CM. If Unified CM is clustered, the fully qualified domain name (FQDN) of each member of the cluster should be listed in the neighbor zone profile of Expressway-C.

A meshed trunk configuration is created here as well. Unified CM will check the status of the nodes in the trunk configuration via a SIP OPTIONS Ping. If a node is not available, Unified CM will take that node out of service and will not route calls to it. Expressway-C will also check the status of the trunk from Unified CM via a SIP OPTIONS Ping. Calls will be routed only to nodes that are shown as active and available. This provides high availability for both sides of the trunk configuration.

DNS SRV records can add to availability of Expressway-E for inbound business-to-business traffic. For high availability, all nodes in the cluster should be listed with the same priority and weight in the SRV record. This allows all nodes to be returned in the DNS query. A DNS SRV record helps to minimize the time spent by a client on lookups because a DNS response can contain all of the nodes listed in the SRV record. The far-end server or far-end endpoint will typically cache the DNS response and will try all nodes returned in the DNS query until a response is received. This provides the best chance for a successful call.

In addition, Expressway clusters support rich media license sharing across clusters. If a node is lost from the cluster, its call licenses will continue to be shared for the next 2 weeks.

Any one Expressway cannot process any more rich media licenses than its physical capacity, even though it can carry more licenses than its physical capacity.

Security for Expressway-C and Expressway-E

Security on Expressway-C and Expressway-E can be further partitioned into network level and application level. Network level security includes feature such as firewall rules and intrusion protection, while application level security includes authorization, authentication, and encryption.

Network Level Protection

Network level protection on Expressway-C and Expressway-E consists of two main components: firewall rules and intrusion protection. Firewall rules enable the ability to:

- Specify the source IP address subnet from which to allow or deny traffic.
- Choose whether to drop or reject denied traffic.
- Configure well known services such as SSH and HTTP/HTTPS, or specify customized rules based on transport protocols and port ranges.
- Configure different rules for the LAN 1 and LAN 2 interfaces on Expressway-E.

The Automated Intrusion Protection feature should be used to detect and block malicious traffic and to help protect the Expressway from dictionary-based attempts to breach login security. Automated Intrusion Protection works by parsing the system log files to detect repeated failures to access specific service categories such as SIP, SSH, and web/HTTPS. When the number of failures within a specified time reaches the configured threshold, the source host IP address (the intruder) and destination port are blocked for a specified period of time. The host address is automatically unblocked after that time period so as not to lock out any genuine hosts that might have been temporarily mis-configured.

Application Level Security

Application level security can be partitioned into:

- Authentication and Encryption, page 5-23
- Dial Plan Protection and Toll Fraud Mitigation, page 5-24

Authentication and Encryption

Securing business-to-business communications includes authentication, encryption, and authorization. Business-to-business communications use an authenticated traversal link by default. The traversal link can also benefit from the use of a Public Key Infrastructure (PKI) verified by a mutually authenticated transport layer security (MTLS) connection between Expressway-C and Expressway-E. If the business-to-business traversal link is deployed on the same Expressway-C and Expressway-E infrastructure as mobile and remote access, make sure that the traversal zone uses the FQDNs of the cluster nodes of Expressway-C and Expressway-E. This makes it straightforward to use certificates for each server to validate the offered certificate against its certificate trust for the traversal connection.

Signaling and media encryption is important for business-to-business calls, but it needs to be deployed carefully so as not to restrict or limit the ability to receive calls. There is a variety of older SIP and H.323 systems that you may be communicating with that do not support signaling or media encryption.

Based on zone configuration, encryption policies might be set as forced (**force encrypted**), desirable (**best effort**), not allowed (**force unencrypted**), or left to the endpoint decisions (**auto**).

If **force encrypted** is configured on a target zone and the Expressway is receiving a call for an endpoint on that remote zone, then Expressway will set up an encrypted call. If the remote party accepts only unencrypted calls, the call will be dropped. If the calling endpoint is using TCP and sending unencrypted media, and **force encrypted** is configured on the target zone, Expressway will terminate the call leg and set up another call leg to the destination with TLS and encryption.

When Expressway performs RTP to SRTP, it uses a back-to-back user agent (B2BUA) for business-to-business calls. The B2BUA terminates both signaling and media and sets up a new call leg to the destination. The B2BUA is engaged any time the media encryption mode is configured to a setting other than **auto**. Note that SIP TLS to TCP interworking requires the B2BUA only if the media is sent encrypted; otherwise it does not require the B2BUA. Exception occurs only in the following scenario affecting Expressway-E: if the inbound zone and outbound zone are set to the same encryption media type and one of those zones is a Traversal Server zone, Expressway-E checks the value of the associated Traversal Client zone. If all three of these zones are set to the same value, the Expressway-E will not engage the B2BUA. In this case, B2BUA will be engaged only on Expressway-C. With **best effort**, if Expressway cannot set up an encrypted call, it will fall back to unencrypted.

Depending on the requirements, different media encryption policies might be configured. If a corporate enforcing policy is not in place, the recommendation is to set up zones with **auto** specified as the media encrypted mode. A setting of **auto** delegates the encryption decisions to endpoints, and Expressway does not perform any sort of RTP-to-SRTP conversion.

When the encryption policy is enforced on Expressway, the call will be divided into many call legs due to B2BUA engagement, as in the following scenario:

- Expressway-C neighbor zone to Unified CM set to **auto**
- Expressway-C traversal client zone set to **best effort**
- Expressway-E traversal server zone set to **best effort**
- Expressway-E DNS zone set to **auto**
- Calling endpoint on Unified CM configured for encryption, and Unified CM configured in mixed mode
- Called endpoint or system does not support encryption

For example, consider a scenario where Unified CM is in mixed mode and the calling endpoint is configured for encryption. In this scenario, a secure endpoint on Unified CM calls an unencrypted endpoint on the Internet. The call will consist of the following call legs:

1. Unified CM endpoint to Expressway-C B2BUA, encrypted
2. Expressway-C B2BUA to Expressway-E B2BUA, encrypted
3. Expressway-E B2BUA to the Internet, up to unknown remote edge or final destination, unencrypted
4. Remote edge to final destination, encrypted or unencrypted depending on called partner's settings

If call legs 1 through 3 are encrypted, the lock icon will display correctly. If one of these legs is not encrypted, the lock icon will not display. Note that the last call leg is under the control of another company, and as such does not influence the lock status.

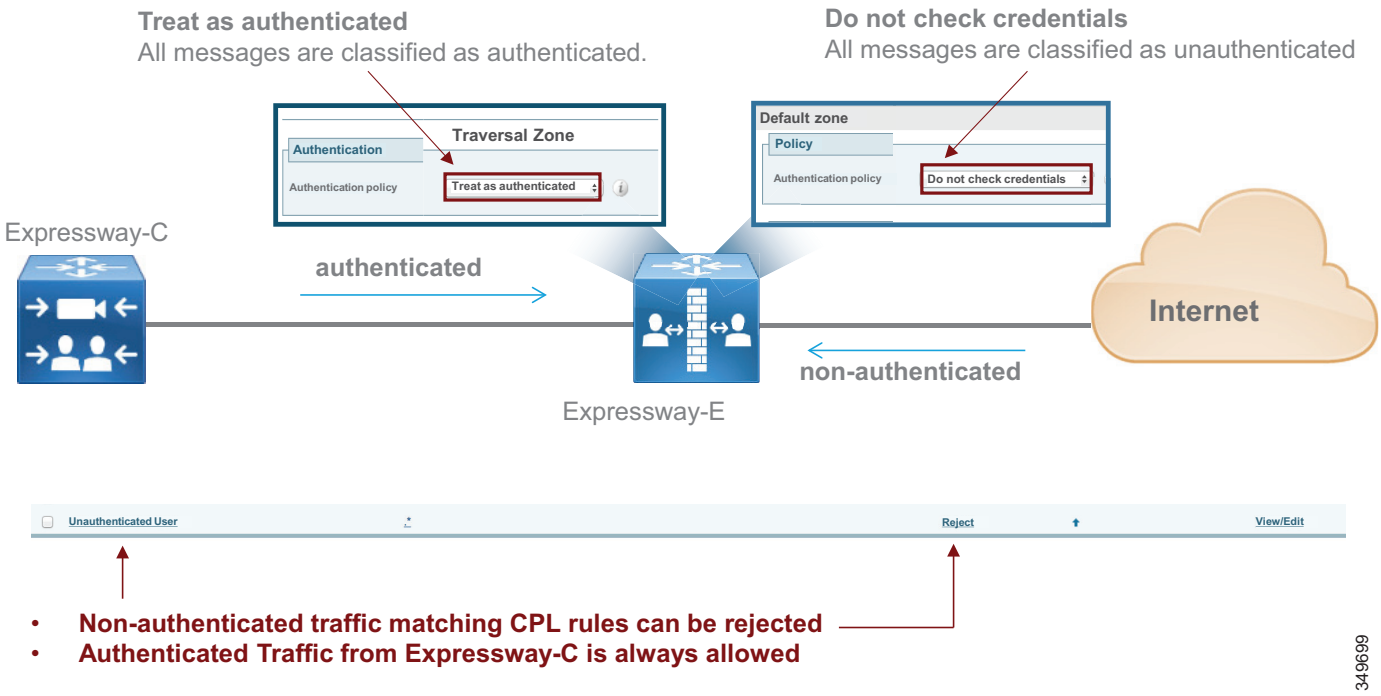
Every company has the control of encryption up to the other company's edge, thus allowing an endpoint to establish an encrypted call from the endpoint to the remote edge. Encryption policy can protect media on the Internet if **force encrypted** is configured on Expressway; but once the call hits the remote edge, the call might be decrypted at the edge level before sending it to the called endpoint.

Dial Plan Protection and Toll Fraud Mitigation

In order to block legal call attempts from unwanted users on the Internet, spam calls, and SIP or H.323 scans, Call Processing Language (CPL) rules can be used on Expressway-E. CPL rules can be applied to call attempts coming from the Internet only.

In order to do this, traffic coming from the traversal client zone can be set to **authenticated**, and traffic from the Internet can be set to **non-authenticated**. CPL rules can be applied to non-authenticated traffic only, bypassing checks for traffic from the internal network or from trusted neighbors on the Internet. Figure 5-9 illustrates this.

Figure 5-9 Zone Authentication Policy



CPL rules are processed using a top-down approach. Two sets of policies can be created:

- Allow-based policy

  An allow-based policy applies regular expressions (regex) to CPL in order allow calls only if they match the numeric range or the alphanumeric URI format internally configured. The last CPL rule will block all calls.

- Deny-based policy

  A deny-based policy denies calls to specific services such as gateways and voicemail, while allowing all the rest if the domain matches the corporate domain. A default CPL rule that blocks all calls is set as last rule.

As an example of a deny-based policy, consider a company where calls are allowed to a set of devices in the range 80XXXXXX only, and where gateway access and other services from external Internet destinations, here represented with 0 and 9, are forbidden. In this case the rules can be set as shown in Table 5-3.

Table 5-3 Example of Deny-Based Policy

Source Type	Destination	Action
Default Zone	**8[1-9]\d{6}@example\.com**	Reject
Default Zone	**[09]\d*@example\.com**	Reject
Default Zone	**\+\d*@example\.com**	Reject
Default Zone	**.*@example\.com**	Allow
Default Zone	**.***	Reject

In addition, it is possible to reject calls based on the calling ID. Unlike the PSTN, where Telecom providers preserve the calling numbers, the Internet is free and nobody is checking the identity of a user. Therefore, it is possible to reject incoming business-to-business calls if the calling alias contains the corporate domain or the IP address of the Expressway-E.

The example in Table 5-4 is based on Cisco Expressway release 8.9.

Table 5-4 Example of Deny-Based Policy Using Expressway-E IP Address

Source Type	Source Alias	Destination Alias	Action
Unauthenticated	**.*@10\.10\.10\.10.***	**.***	Reject
Unauthenticated	**.*@example\.com.***	**.***	Reject

10.10.10.10 in Table 5-4 represents the public address of Expressway-E. These rules can be added to the previous list just before the "allow" rule. In this way any call from the Internet containing the corporate domain or IP address will be rejected, thus mitigating identity theft.

Because the Default Zone is the target for business-to-business incoming calls, it has to be configured with an authentication policy set to "do not check credentials." In this way business-to-business calls will be considered unauthenticated and thus checked against the rules. Internal traffic coming from the Traversal Zone will bypass this check if that zone is configured with an authentication policy set to "treat as authenticated", as shown in Figure 5-9.

Scaling the Expressway Solution

When multiple Internet edges are deployed, it is important to set routing rules properly in order to send collaboration traffic to the nearest Internet edge.

Multiple Expressway-Es and GeoDNS

Scalability for business-to-business communications can be addressed by adding multiple Expressway-C and Expressway-E clusters, either in the same physical location or geographically dispersed. When multiple Expressway-C and Expressway-E pairs are deployed, Unified CM can direct an outbound call to the edge server that is nearest to the calling endpoint, thus minimizing internal WAN traffic. For large deployments it might be preferable to host business-to-business communications on Expressway-C and Expressway-E pairs separate from mobile and remote access. This allows the server resources to be dedicated to external Internet communications.

When two or more Internet edges are deployed, it is important to understand how to split the load between them. If the Internet edges are deployed in the same data center or in the same area, load balancing can occur at the DNS SRV level. As an example, if the enterprise network includes three Internet edges used for business-to-business communications, each one consisting of a cluster of two Expressway-E and Expressway-C nodes, the _sips._tcp.example.com and _sip._tcp.example.com records will include all six Expressway-E records at the same priority and weight. This distributes the registrations and calls equally across the various Expressway-E and Expressway-C clusters.

However, if the Expressway clusters are deployed across geographical regions, some intelligent mechanisms on top of the DNS SRV priority and weight record are needed to ensure that the endpoint uses the nearest Expressway-E cluster. As an example, if an enterprise has two Expressway clusters, one in the United States (US) and the other in Europe (EMEA), it is desirable for users located in the US to be directed to the Expressway-E cluster in the US while users in Europe are directed to the Expressway-E cluster in Europe. This is facilitated by implementing GeoDNS services. GeoDNS services are cost effective and easy to configure. With GeoDNS it is possible to route traffic based on different policies such as location (IP address routing), minimum latency, and others.

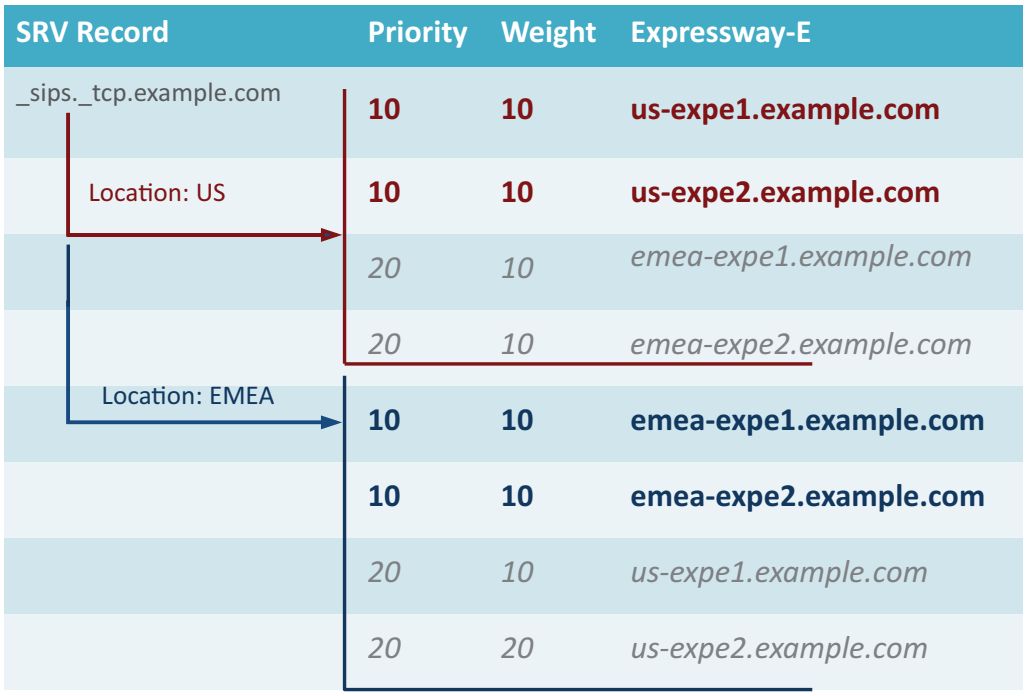
The following examples explain how to configure DNS for GeoDNS services.

In our example scenario, two Internet edge Expressway clusters are deployed, one in the US and one in Europe, each composed of two Expressway-C and Expressway-E servers. If the measured latency between the calling endpoint and the European edge is less than the latency between the endpoint and the US edge, or if the endpoint IP address matches the range for the US, the endpoint will be directed to the European edge for registration based on the configured policy (latency or IP address).

Although some GeoDNS providers support GeoDNS services on SRV records, many others allow GeoDNS for CNAME or A-records only. The recommendation is to implement GeoDNS services on SRV records because this allows for a simpler configuration and easy troubleshooting. A GeoDNS configuration for SRV records is shown in the following example.

If the calling user is in the US, the call will be sent to the US; but if the US data center is down, the call will be sent to EMEA. This configuration allows for geographic redundancy and is shown in Figure 5-10.

Figure 5-10 GeoDNS Configuration for SRV Records



However, if your GeoDNS provider allows you to specify GeoDNS services for CNAME records only and not for SRV records, the following example shows how to configure the GeoDNS if only CNAME is supported for GeoDNS services.

Following this scenario, a DNS SRV record resolves into a CNAME record which, in turn, resolves into an A-record. CNAME records can be assigned a geographic location. As an example, consider an Expressway-E cluster in the US and another Expressway-E cluster in EMEA. A SRV record *_sips._tcp.example.com* for SIP TLS and/or *_sip._tcp.example.com* is configured for business-to-business calls. This record resolves into *alias1.example.com*, a CNAME record.

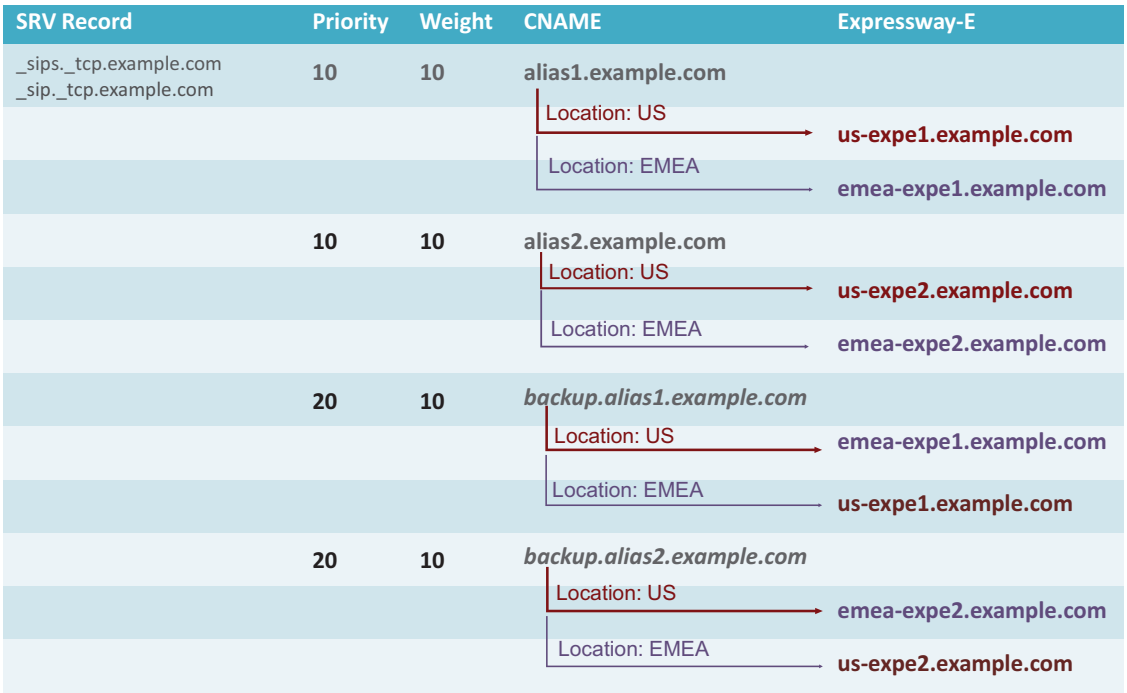
Based on the GeoDNS configuration, a label is applied to the CNAME record to identify the region where the record is active. In this case, the CNAME resolution will be an A-record for the US and another A-record for EMEA, with highest priority (10 in this example). This will address the first peer of the cluster in both regions.

The second CNAME record will be resolved into the second peer of US and EMEA clusters with highest priority. This needs to be repeated until all peers of the cluster are included.

In order to have geographic redundancy, backup CNAME aliases have to be created. In the example in Figure 5-11, *backup-alias1.example.com* resolves into the first EMEA Expressway peer for US users and into the first US Expressway peer for EMEA users, thus providing geographic redundancy for both regions. This backup alias process has to be repeated until all peers of the cluster are included. Those backup records will be used only if the first ones are not answering, because the DSN SRV is set to a lower priority (20 in the example).

Figure 5-11 shows the DNS record structure for GeoDNS services applied to CNAME records.

Figure 5-11 GeoDNS Record Structure for CNAME Records with Geographic Redundancy



Two Different Expressway Edges without GeoDNS

Though the recommendation is to adopt the GeoDNS approach, there might be cases where GeoDNS cannot be deployed; for example, in those cases where the DNS records are managed by a service provider that does not offer the GeoDNS services, or when the multiple edges are deployed in regions that are smaller than the capacity of GeoDNS to select between them. As an example, GeoDNS might be able to distinguish if the calling endpoint location is in California or in Pennsylvania, but it might be not be able to distinguish if the calling location endpoint is San Jose or San Diego. So GeoDNS could not be used if the two Expressway clusters are located in San Jose and in San Diego.

An alternative solution is designed to return the edge that is closest to the destination endpoint or device. This requires finding or knowing where the destination endpoint is located and then returning the appropriate edge. The benefit of this solution is to minimize the use of bandwidth on the customer network by delivering the shortest internal path to the endpoint.

In this scenario, business-to-business SRV records are set with the same priority and weight for all Expressway servers.

As an example, consider two Expressway-C and Expressway-E clusters in EMEA, and another two Expressway-C and Expressway-E clusters in APJC. The Unified CM inbound calling search space on the Expressway-C trunk in EMEA will contain the partition of the EMEA phones but not the partition of the APJC phones. Analogously, the inbound calling search space on the Expressway-C trunk in APJC will contain the partition of the APJC phones but not the partition of the EMEA phones. If a user on the Internet in EMEA calls a corporate endpoint in APJC, the call might be sent to either the EMEA or APJC Expressway cluster.

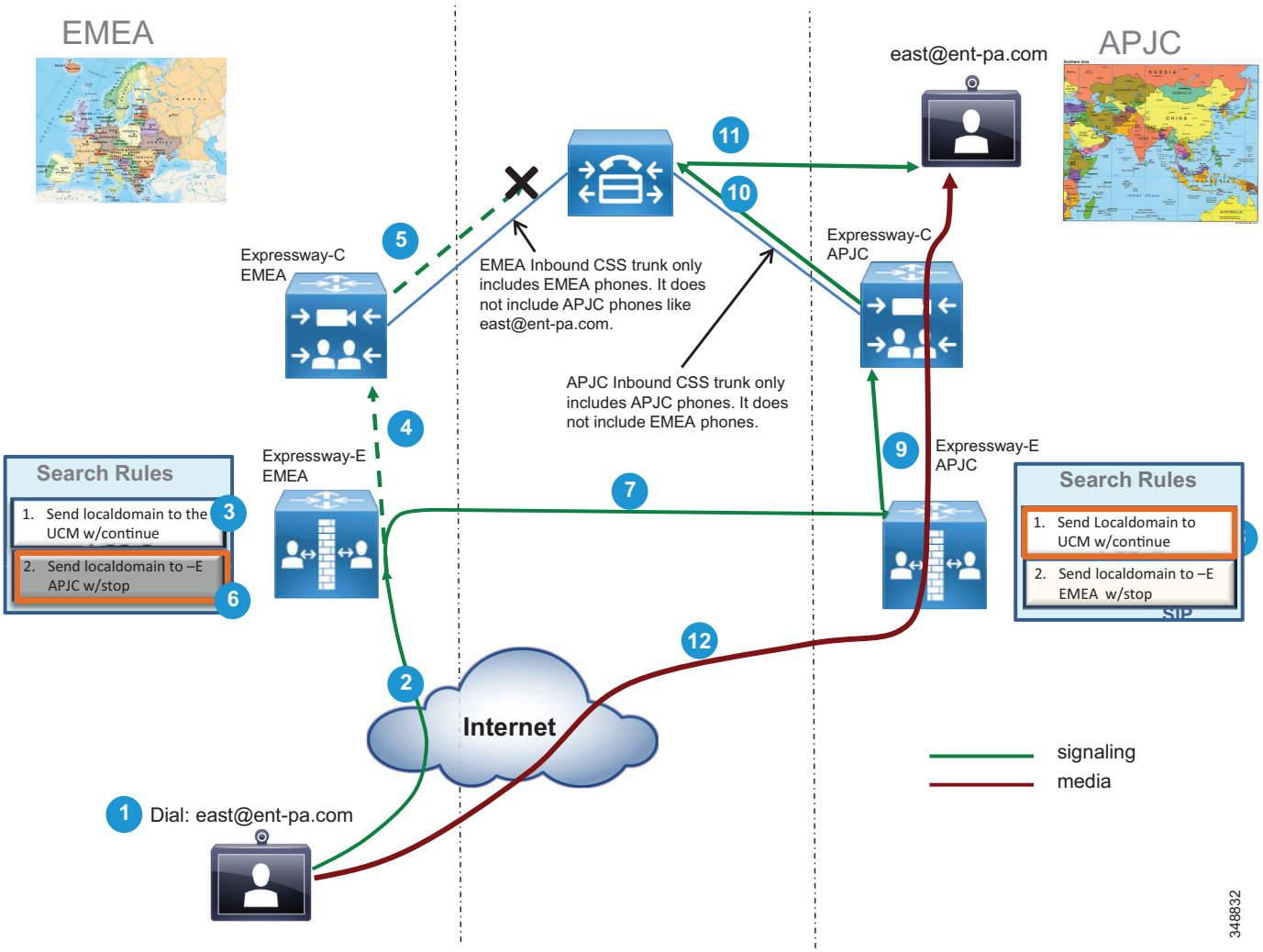
In this example, assume that the call is sent to the EMEA Expressway-E cluster. The EMEA Expressway-E and Expressway-C will try to send the call to the destination, but the inbound calling search space of the Expressway-C trunk will block the call. The EMEA Expressway-E will then forward the call to the APJC Expressway-E. This time the call will be delivered to the destination because the inbound calling search space of APJC Expressway-C contains the APJC endpoint's partition.

In order to allow the Expressway-E in EMEA to remove itself from the signaling and media path, it is important to make sure that there is no TCP-to-TLS or RTP-to-SRTP conversion on Expressway-E EMEA clusters, and to make sure that the call signaling optimization parameter is set to **On** in all Expressway-C and Expressway-E nodes.

Because this is not a deterministic process, in the case of three or more Expressway edges the searching mechanism would require too much time. Therefore, this configuration is recommended for no more than two Expressway edges. If more than two edges are required, the recommendation is to deploy a Directory Expressway architecture. Directory Expressway architecture is not covered in this document.

Figure 5-12 shows the Expressway edge design that enables selection of the edge closest to the destination endpoint.

Figure 5-12 Selection of the Expressway Cluster Closest to the Destination

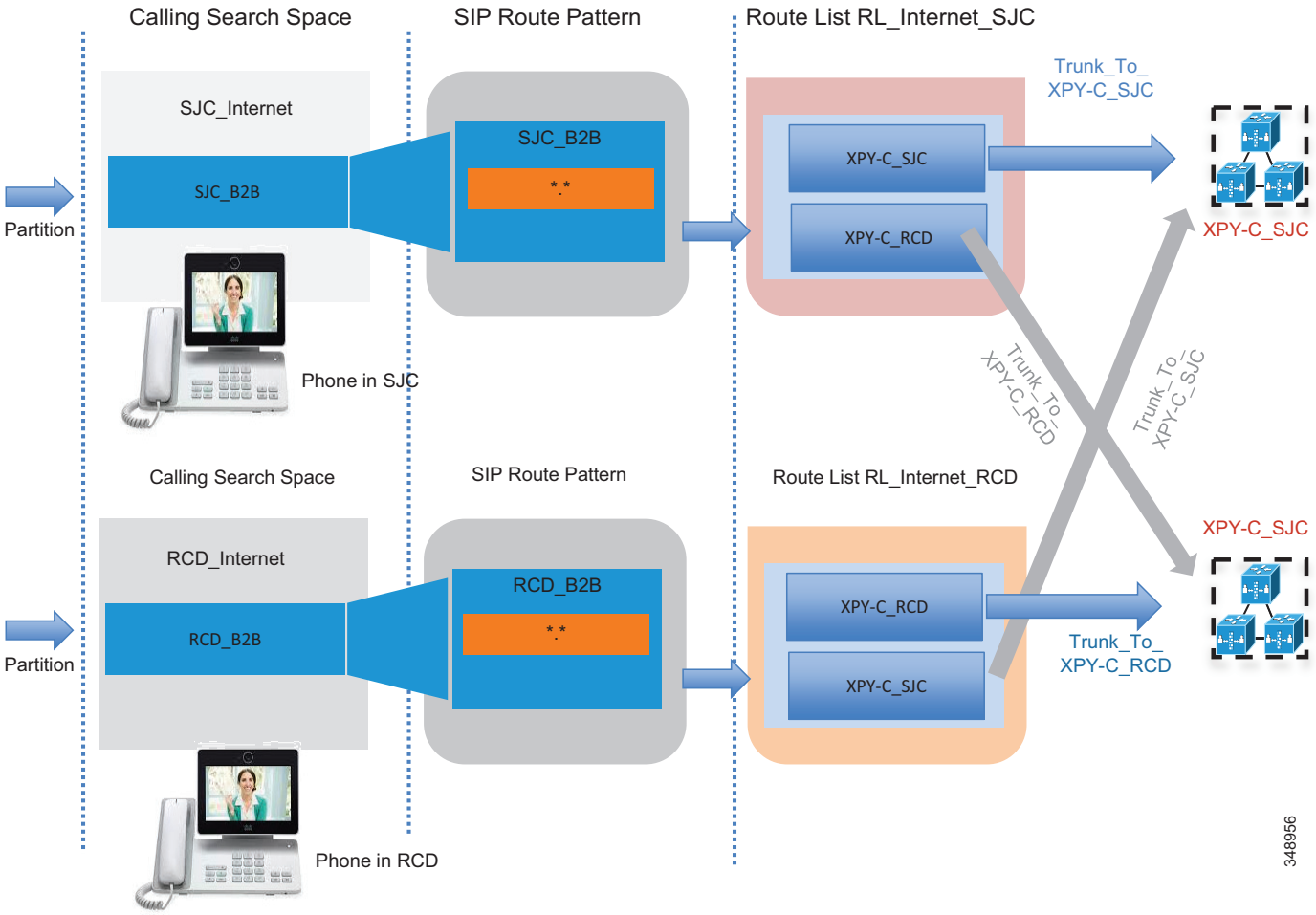


This architecture can scale to more than two sites, and it needs a central Expressway node called Directory Expressway. Directory Expressway is an Expressway acting as a transit node between Expressways in different regions. Directory Expressway architecture is not currently covered in this document.

Considerations for Outbound Calls

Outbound calls should be directed to the Expressway-C that is nearest to the calling endpoint. This can be achieved by using Cisco Unified CM mechanisms such as calling search spaces and partitions. Figure 5-13 shows the Unified CM configuration.

Figure 5-13 Unified CM Configuration to Direct Outbound Calls to the Nearest Expressway-C Cluster, and Use a Backup Cluster if the Nearest One is not Available



The Unified CM Local Route Group feature helps scale this solution when multiple sites access two or more Expressway-C clusters. This mechanism is also applied on ISDN gateways and Cisco Unified Border Element. A full description of the configuration is documented in the next two sections, since it also applies to Cisco Unified Border Element and voice gateways.

Best Practices for Gateways

This section addresses the following best practices with regard to gateways:

- Tuning Gateway Gain Settings, page 5-32
- Routing Inbound Calls from the PSTN, page 5-32
- Routing Outbound Calls to the PSTN, page 5-33
- Automated Alternate Routing (AAR), page 5-34
- Least-Cost Routing, page 5-36

Tuning Gateway Gain Settings

Connecting a Cisco Unified Communications network to the PSTN through gateways requires that you properly address media quality issues arising from echo and signal degradation due to power loss, impedance mismatches, delay, and so forth. For this purpose, you must establish a Network Transmission Loss Plan (NTLP), which provides a complete picture of signal loss in all expected voice paths. Using this plan, you can identify locations where signal strength must be adjusted for optimum loudness and effective echo cancellation. Note that not all carriers use the same loss plan, and that the presence of cellular networks adds further complexity in creating the NTLP. Cisco does not recommend adjusting input gain and output attenuation on gateways without first completing such an NTLP. For more information, refer to *Echo Analysis for Voice Over IP*, available at

https://www.cisco.com/en/US/docs/ios/solutions_docs/voip_solutions/EA_ISD.pdf

Routing Inbound Calls from the PSTN

Use one of the following methods to route inbound calls from the PSTN:

- Assign a single directory number to each user for both video and voice calls. This method is not recommended because all calls would have to be received from the PSTN on a video gateway, including audio-only calls. This would waste valuable video gateway resources and be hard to scale.
- Assign at least two different directory numbers to each video-enabled device in the Unified CM cluster, with one line for audio and another line for video. With this method, the outside (PSTN) caller must dial the correct number to enable video.
- For video calls, have outside callers dial the main number of the video gateway. Cisco ISDN and Serial gateways offer an integrated auto-attendant that prompts the caller to enter the extension number of the party they are trying to reach. Unified CM will then recognize that it is a video call when ringing the destination device. This method relieves the caller from having to remember two different DID numbers for each called party, but it adds an extra step to dialing an inbound video call.

Note The outside video endpoints must support DTMF in order to enter the extension of the called party at the IVR prompt.

The following example illustrates the second method:

A user has a Cisco Unified IP Phone with video capabilities enabled. The extension of the IP Phone is 51212, and the fully qualified DID number is 1-408-555-1212. To reach the user from the PSTN for a voice-only call, people simply dial the DID number. The CO sends calls to that DID number through T1-PRI circuit(s) connected to a Cisco Voice Gateway. When the call is received by the gateway, Unified CM knows that the gateway is capable of audio only, so it negotiates only a single audio channel for that call. Conversely, for people to reach the user from the PSTN for a video call, they must dial the main number of the video gateway and then enter the user's extension. For example, they might dial 1-408-555-1000. The CO would send calls to that number through the T1-PRI circuit(s) connected to a Cisco ISDN video gateway. When the call is received by the gateway, an auto-attendant prompt asks the caller to enter the extension of the person they are trying to reach. When the caller enters the extension via DTMF tones, Unified CM knows that the gateway is capable of video, so it negotiates both audio and video channels for that call.

Gateway Digit Manipulation

The Cisco TelePresence ISDN Gateways 8321 and 3241 and the Cisco TelePresence Serial Gateways 8330 and 3340 all have capabilities for digit manipulation. It is possible to set up multiple dial plan rules on these video gateways. These rules match based on calling and/or called number and work in either the IP-to-PSTN or PSTN-to-IP direction. When an inbound call matches a configured dial plan rule, the ISDN or Serial gateway can take one of the following actions:

- Reject the call
- Enter the Auto Attendant
- Place a call to a number (or IP address, hostname, or URI in the case of a PSTN-to-IP call)

When the action is to place a call to a number, the original called number or parts of it can be used in the new number to call.

For more details, refer to the following documentation:

- Cisco TelePresence ISDN Gateway documentation, available at

  https://www.cisco.com/en/US/products/ps11448/tsd_products_support_series_home.html
- Cisco TelePresence Serial Gateway documentation, available at

  https://www.cisco.com/en/US/products/ps11605/tsd_products_support_series_home.html

Routing Outbound Calls to the PSTN

Use one of the following methods to route outbound calls to the PSTN:

- Assign different access codes (that is, different route patterns) for voice and video calls. For example, when the user dials 9 followed by the PSTN telephone number they are trying to reach, it could match a route pattern that directs the call out a voice gateway. Similarly, the digit 8 could be used for the route pattern that directs calls out a video gateway.
- Assign at least two different directory numbers on each video-enabled device in the Unified CM cluster, with one line for audio and another line for video. The two lines can then be given different calling search spaces. When users dial the access code (9, for example) on the first line, it could be directed out a voice gateway, while dialing the same access code on the second line could direct the call out a video gateway. This method alleviates the need for users to remember two different access

codes but requires them to press the correct line on their phones when placing calls. However not all Cisco video endpoints support multiple lines at this time, in which case prefixes would be the preferred method for routing outbound calls to the PSTN.

Video Gateway Call Bandwidth

The Cisco TelePresence ISDN Gateway dial plan rules can be configured so that calls with a certain prefix are limited to a maximum amount of bandwidth on the ISDN connection to the PSTN. This is useful to ensure that a single call cannot monopolize the entire PRI link. When you configure a service prefix in the gateway, you can choose one of the following maximum speeds:

- 128 kbps
- 192 kbps
- 256 kbps
- 320 kbps
- 384 kbps
- 512 kbps
- 768 kbps
- 1152 kbps
- 1472 kbps

Calls from an IP endpoint toward the PSTN can include the service prefix at the beginning of the called number in order for the gateway to decide which service to use for the call. Optionally, you can configure the default prefix to be used for calls that do not include a service prefix at the beginning of the number. This method can become quite complex because users will have to remember which prefix to dial for the speed of the call they wish to make, and you would have to configure multiple route patterns in Unified CM (one for each speed).

Note Two global settings on the Cisco TelePresence ISDN Gateway can be used to set a minimum or maximum bandwidth value for incoming and outgoing ISDN calls. The dial plan cannot override this value with a higher maximum bandwidth; however, a dial plan can impose a lower bandwidth for particular calls.

Automated Alternate Routing (AAR)

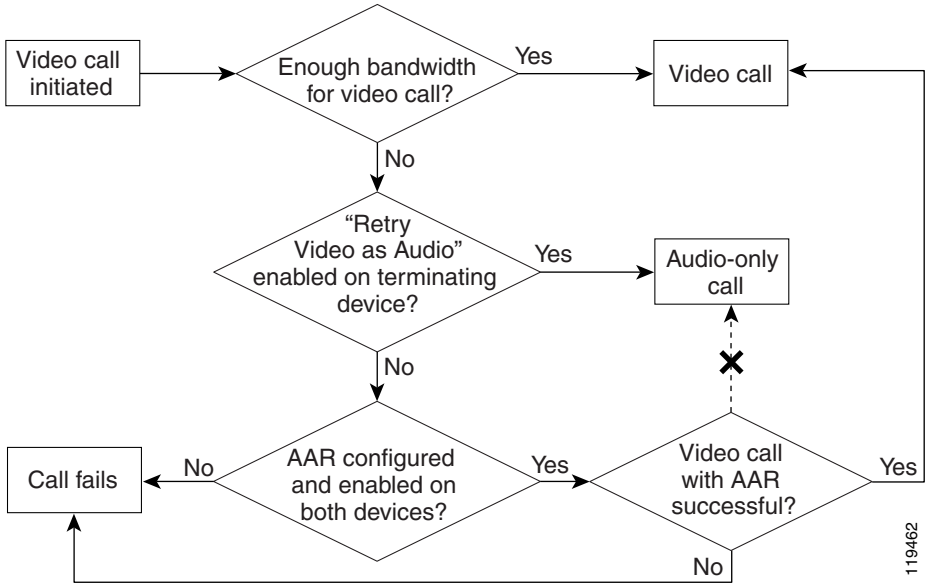
When the IP network does not have enough bandwidth available to process a call, Unified CM uses its call admission control mechanism to determine what to do with the call. Unified CM performs one of the following actions with the call, depending on how you have configured it:

- Fail the call, playing busy tone to the caller and displaying a Bandwidth Unavailable message on the caller's screen
- Retry a video call as an audio-only call
- Use automated alternate routing (AAR) to re-route the call over an alternative path, such as a PSTN gateway

The Retry Video Call as Audio option takes effect only on the terminating (called) device, thus allowing the flexibility for the calling device to have different options (retry or AAR) for different destinations.

If a video call fails due to bandwidth limitations but automated alternate routing (AAR) is enabled, Unified CM will attempt to reroute the failed call as a video call to the AAR destination. If AAR is not enabled, the failed call will result in a busy tone and an error message will be sent to the caller. (See Figure 5-14.)

Figure 5-14 Possible Scenarios for a Video Call



To provide AAR for voice or video calls, you must configure the calling and called devices as members of an AAR group and configure an External Phone Number Mask for the called device. The External Phone Number Mask designates the fully qualified E.164 address for the user's extension, and the AAR group indicates what digits should be prepended to the External Phone Number Mask of the called device in order for the call to route successfully over the PSTN.

For example, assume that user A is in the San Jose AAR group and user B is in the San Francisco AAR group. User B's extension is 51212, and the External Phone Number Mask is 6505551212. The AAR groups are configured to prepend 91 for calls between the San Jose and San Francisco AAR groups. Thus, if user A dials 51212 and there is not enough bandwidth available to process the call over the IP WAN between those two sites, Unified CM will take user B's External Phone Number Mask of 6505551212, prepend 91 to it, and generate a new call to 916505551212 using the AAR calling search space for user A.

By default, all video-capable devices in Unified CM have the Retry Video Call as Audio option enabled (checked). Therefore, to provide AAR for video calls, you must disable (uncheck) the Retry Video Call as Audio option. Additionally, if a call admission control policy based on Resource Reservation Protocol (RSVP) is being used between locations, the RSVP policy must be set to Mandatory for both the audio and video streams.

Furthermore, Unified CM looks at only the called device to determine whether the Retry Video Call as Audio option is enabled or disabled. So in the scenario above, user B's phone would have to have the Retry Video Call as Audio option disabled in order for the AAR process to take place.

Finally, devices can belong to only one AAR group. Because the AAR groups determine which digits to prepend, AAR groups also influence which gateway will be used for the rerouted call. Depending on your choice of configuration for outbound call routing to the PSTN, as discussed in the previous section,

video calls that are rerouted by AAR might go out a voice gateway instead of a video gateway. Therefore, carefully construct the AAR groups and the AAR calling search spaces to ensure that the correct digits are prepended and that the correct calling search space is used for AAR calls.

While these considerations can make AAR quite complex to configure in a large enterprise environment, AAR is easier to implement when the endpoints are strictly of one type or the other. When endpoints are capable of both audio and video calls (such as Cisco Unified IP Phone 9971 or a Cisco TelePresence System EX90), the configuration of AAR can quickly become unwieldy. Therefore, Cisco recommends that large enterprise customers who have a mixture of voice and video endpoints give careful thought to the importance of AAR for each user, and use AAR only for select video devices such as dedicated videoconference rooms or executive video systems. Table 5-5 lists scenarios when it is appropriate to use AAR with various device types.

Table 5-5 When to Use AAR with a Particular Device Type

Device Type	Device is used to call:	Enable AAR?	Comments
IP Phone	Other IP Phones and video-capable devices	Yes	Even when calling a video-capable device, the source device is capable of audio-only, thus AAR can be configured to route calls out a voice gateway.
Cisco Jabber or Cisco Unified IP Phone 9971	Other video-capable devices only	Yes	Because the device is used strictly for video calls, you can configure the AAR groups accordingly.
	IP Phones and other video-capable devices	No	It will be difficult to configure the AAR groups to route audio-only calls differently than video calls.
H.323 or SIP client	Other video-capable devices only	Yes	Because the device is used strictly for video calls, you can configure the AAR groups accordingly.
	IP Phones and other video-capable devices	No	It will be difficult to configure the AAR groups to route audio-only calls differently than video calls

Least-Cost Routing

Least-cost routing (LCR) and tail-end hop-off (TEHO) are very popular in VoIP networks and can be used successfully for video calls as well. In general, both terms refer to a way of configuring the call routing rules so that calls to a long-distance number are routed over the IP network to the gateway closest to the destination, in an effort to reduce toll charges. Unified CM supports this feature through its rich set of digit analysis and digit manipulation capabilities, including:

- Partitions and calling search spaces
- Translation patterns
- Route patterns and route filters
- Route lists and route groups

Configuring LCR for video calls is somewhat more complicated than for voice calls, for the following reasons:

- Video calls require their own dedicated gateways, as discussed previously in this chapter
- Video calls require much more bandwidth than voice calls

With respect to dedicated gateways, the logic behind why you might or might not decide to use LCR for video calls is very similar to that explained in the section on Automated Alternate Routing (AAR), page 5-34. Due to the need to have different types of gateways for voice and video, it can become quite complex to configure all the necessary partitions, calling search spaces, translation patterns, route patterns, route filters, route lists, and route groups needed for LCR to route voice calls out one gateway and video calls out another.

With respect to bandwidth requirements, the decision to use LCR depends on whether or not you have enough available bandwidth on your IP network to support LCR for video calls to/from a given location. If the current bandwidth is not sufficient, then you have to determine whether the benefits of video calls are worth the cost of either upgrading your IP network to make room for video calls or deploying local gateways and routing calls over the PSTN. For example, suppose you have a central site with a branch office connected to it via a 1.544-Mbps T1 circuit. The branch office has twenty video-enabled users in it. A 1.544-Mbps T1 circuit can handle at most about four 384-kbps video calls. Would it really make sense in this case to route video calls up to the central site in order to save on toll charges? Depending on the number of calls you want to support, you might have to upgrade your 1.544-Mbps T1 circuit to something faster. Is video an important enough application to justify the additional monthly charges for this upgrade? If not, it might make more sense to deploy a Cisco video gateway at the branch office and not bother with LCR. However, placing local Cisco video gateways at each branch office is not inexpensive either, so ultimately you must decide how important video-to-PSTN calls are to your business. If video is not critical, perhaps it is not worth upgrading the bandwidth or buying video gateways but, instead, using the Retry Video Call as Audio feature to reroute video calls as voice-only calls if they exceed the available bandwidth. Once a call is downgraded to voice-only, local gateway resources and bandwidth to perform LCR become more affordable and easier to configure.

Fax and Modem Support

For information on fax and modem support across Cisco gateways refer to the following documentation:

- The *Gateways* chapter of the *Cisco Unified Communications System 9.0 SRND*, available at https://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/9x/gateways.html
- *Fax, Modem, and Text Support over IP Configuration Guide*, available at https://www.cisco.com/en/US/docs/ios-xml/ios/voice/fax/configuration/15-mt/vf-15-mt-book.html

Cisco Unified CM Trunks

Revised: February 7, 2017

A trunk is a communications channel on Cisco Unified Communications Manager (Unified CM) that enables Unified CM to connect to other servers. Using one or more trunks, Unified CM can receive or place voice, video, and encrypted calls, exchange real-time event information, and communicate in other ways with call control servers and other external servers.

Trunks are an integral and crucial part of a Cisco Collaboration System deployment, hence it is important to understand the types of trunks available, their capabilities, and design and deployment considerations such as resiliency, capacity, load balancing, and so forth.

There are two basic types of trunks that can be configured in Unified CM:

- SIP and H.323 trunks, both of which can be used for external communications
- Intercluster trunks (ICTs)

While H.323 trunks are still supported, SIP trunks are the recommended trunk type for Unified Communication deployments because SIP trunks provide additional features and functionality not available with H.323 trunks. This chapter provides a comparative overview of the capabilities of H.323 and SIP trunks, but the focus of this chapter is on SIP trunks, their operation, and features for Unified Communications deployments. For detailed information on H.323 trunk capabilities and operation, refer to the *Cisco Unified CM Trunks* chapter of the *Cisco Collaboration 9.x SRND*, available at

https://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/collab09/trunks.html

For more details on the applications of Unified CM trunks, refer to their respective sections in the following chapters of this document:

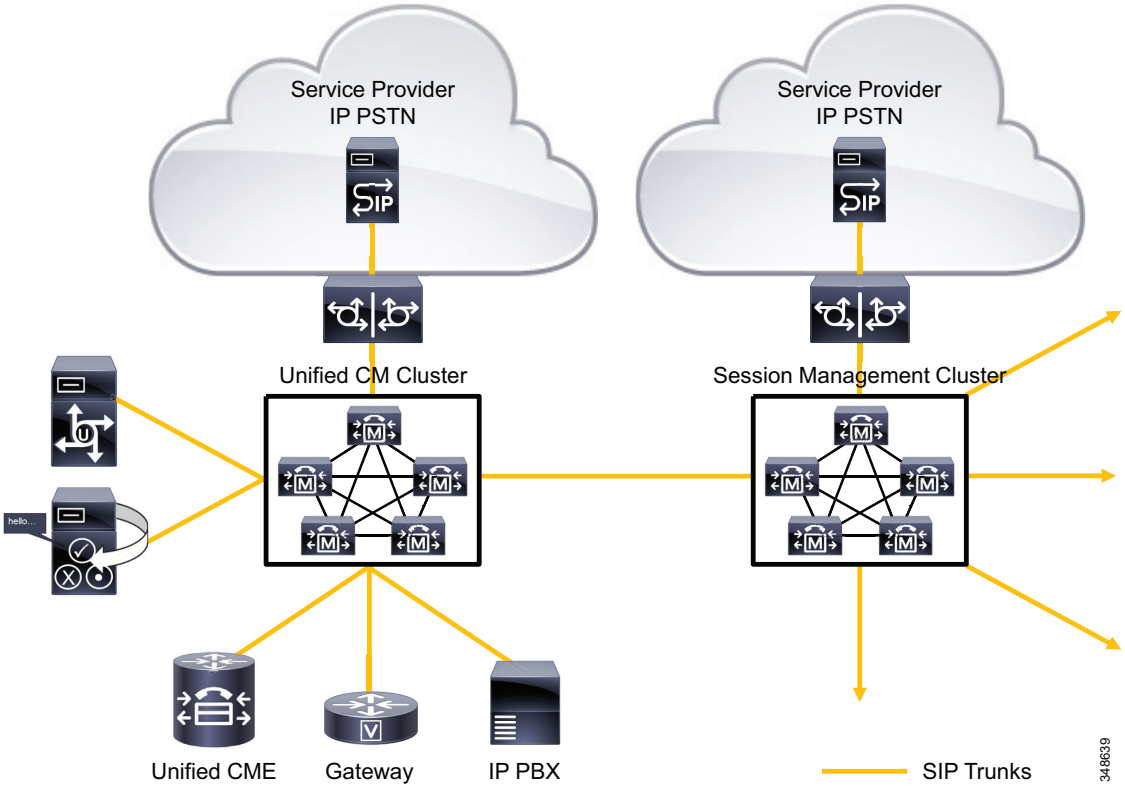
- Collaboration Deployment Models, page 10-1
- Media Resources, page 7-1
- Bandwidth Management, page 13-1
- Collaboration Instant Messaging and Presence, page 20-1

Unified CM Trunks Solution Architecture

Cisco Unified CM uses the mechanism of IP trunks to exchange call-related information with other components of a Unified Communications solution. Given their importance in this respect, it is important to develop the system architecture of the IP trunks with proper regard to the protocol, feature and service expectations, performance requirements, and so forth.

Figure 6-1 illustrates the role of IP trunks in system connectivity. The illustration does not show all possible connections from the Unified CM cluster.

Figure 6-1 IP Trunks Provide Connections to Unified CM



Calls are directed toward trunks as defined by the dial plan, using the route pattern construct. A route pattern can use a trunk either directly or through a route list. The route list, if used, consists of one or more route groups, each of which contains one or more trunks. An individual trunk within a route group may be configured to be selected in either a top-down or circular fashion. For outgoing calls, Unified CM selects one of the trunks associated in this fashion with the route pattern. Before it accepts an incoming call, Unified CM verifies whether a trunk is defined to the remote address from which the call is received.

A Comparison of SIP and H.323 Trunks

Cisco Unified CM trunk connections support both SIP and H.323. The decision to use SIP or H.323 is driven by the unique feature(s) offered by each protocol. Over the past several releases, as SIP has grown in popularity among both Unified Communications vendors and customers, the features and functionality supported by SIP trunks have grown to the point where SIP trunks offer a richer set of features than H.323 trunks, making SIP trunks the recommended choice for Unified Communications deployments. Today, most customers are migrating away from H.323 trunks and gatekeeper-based Unified Communications deployments to those that use SIP trunks only, with Cisco Unified Communications Manager Session Manager Edition as the trunk and dial plan aggregation platform.

As can be seen in Table 6-1, while SIP and H.323 trunks share many of the same features for trunk connections between Cisco devices, SIP trunks support several features that are not supported by H.323 trunks. For trunk connections to other vendors' products and to service provider networks, SIP is the most commonly deployed protocol today and is growing in usage as Unified Communications products and networks using protocols such as H.323 migrate to SIP.

Table 6-1 compares some of the features offered over SIP and H.323 trunks between Unified CM clusters.

Table 6-1 Comparison of SIP and H.323 Features on Cisco Unified CM Trunks

Feature	SIP	H.323
Calling Line (Number) Identification Presentation	Yes	Yes
Calling Line (Number) Identification Restriction	Yes	Yes
Calling Name Identification Presentation	Yes	Yes
Calling Name Identification Restriction	Yes	Yes
Connected Line (Number) Identification Presentation	Yes	Yes
Connected Line (Number) Identification Restriction	Yes	Yes
Connected Name Identification Presentation	Yes	Yes
Connected Name Identification Restriction	Yes	Yes
Alerting Name	Yes	***No***
Call Transfer (Blind/Attended)	Yes/Yes	Yes/Yes
Call Forward All	Yes	Yes
Call Forward Busy	Yes	Yes
Call Forward No Reply	Yes	Yes
QSIG Call Completion to Busy Subscriber	Yes	Yes
QSIG Call Completion No Reply	Yes	Yes
QSIG Path Replacement	Yes	Yes
Subscribe/Notify, Publish - Presence	Yes	***No***
Message Waiting Indication (MWI: lamp ON, lamp OFF)	Yes	***No***
Call Hold/Resume	Yes	Yes
Music On Hold (unicast and multicast)	Yes	Yes

Table 6-1 Comparison of SIP and H.323 Features on Cisco Unified CM Trunks (continued)

Feature	SIP	H.323
DTMF-relay	RFC 2833, KPML (OOB), Unsolicited Notify (OOB)	H.245 Out Of Band (OOB)[1]
SIP Early Offer	Yes - MTP may be required	N/A
Best Effort Early Offer	Yes - No MTPs used. SIP Early Offer sent if possible; if not, SIP Delayed Offer sent.	N/A
SIP Delayed Offer	Yes	N/A
H.323 Fast Start	N/A	Yes - MTP always required for Outbound Fast Start - Voice Calls Only supported
Accept Audio Codec Preference in Received Offer	Yes	***No***
Codecs with MTP for SIP Early Offer/ H323 Fast Start	All codecs supported when **Early Offer support for voice and video calls - Mandatory (insert MTP if needed)** or **Early Offer support for voice and video calls - Best Effort (no MTP inserted)** is selected G.711, G.729 when **MTP Required** is selected	G.711, G.723, G.729 only
Video	Yes	Yes
Video codecs	H.261, H.263, H.263+, H.264 AVC	H.261, H.263, H.263+, H.264 AVC
Video Presentation sharing (BFCP)	Yes	***No***
Multi-Level Precedence and Preemption (MLPP)	Yes	Yes
T.38 Fax	Yes	Yes
Signaling Authentication	Digest, TLS	***No***
Signaling Encryption	TLS	***No***
Media Encryption (audio)	SRTP	SRTP
RSVP-based QoS and call admission control	Yes	***No***
Support for + character	Yes	***No***
Incoming Called Party Transformations	Yes	Yes
Incoming Calling Party Transformations	Yes	Yes

Table 6-1 Comparison of SIP and H.323 Features on Cisco Unified CM Trunks (continued)

Feature	SIP	H.323
Connected Party Transformation	Yes	Yes
Outbound Calling Party Transformations	Yes	Yes
Outbound Called Party Transformations	Yes	Yes
Outbound Calling/Called Party Number Type Setting	SIP does not support Number Type	Unified CM, Unknown, National, International, Subscriber
Outbound Called/Called Party Numbering Plan Setting	SIP does not support Number Plan	Unified CM, ISDN, National Standard, Private, Unknown
Trunk destination - State detection mechanism	OPTIONS Ping	***Per call attempt***
IPv6, Dual Stack, ANAT	Yes	***No***
Protocol modification scripts for interoperability	Yes	***No***
Run on All Unified CM Nodes	Yes	Yes
Up to 16 Destination Addresses	Yes	Yes
URI based calls	Yes	No
Geo Location support	Yes	Yes

1. H.323 trunks support signaling of RFC 2833 for certain connection types.

SIP Trunks Overview

SIP trunks provide connectivity to other SIP devices such as gateways, Cisco Unified CM Session Management Edition, SIP proxies, Unified Communications applications, and other Unified CM clusters. Today, SIP is arguably the most commonly chosen protocol when connecting to service providers and Unified Communications applications. Cisco Unified CM provides the following SIP trunk and call routing features:

- Runs on all Unified CM nodes
- Up to 16 destination IP addresses per trunk
- SIP OPTIONS ping keep-alives
- Early Offer support for voice and video calls Mandatory (insert MTP if needed)
- Early Offer support for voice and video calls Best Effort (No MTP inserted) — also known as Best Effort Early Offer
- Audio codec preference (and Accept Audio Codec Preference in Received Offer)
- SIP trunk normalization and transparency scripts for interoperability
- SIP REFER transparency
- H.264 Video with Desktop Presentation (Binary Floor Control Protocol (BFCP)) and Far End Camera Control (FECC)

The SIP trunk features available in the current release of Unified CM make SIP the preferred choice for new and existing trunk connections. The QSIG over SIP feature provides parity with H.323 intercluster trunks and can also be used to provide QSIG over SIP trunk connections to Cisco IOS gateways (and on

to QSIG-based TDM PBXs). The ability to run on all Unified CM nodes and to handle up to 16 destination IP addresses improves outbound call distribution from Unified CM clusters and reduces the number of SIP trunks required between clusters and devices. SIP OPTIONS ping provides dynamic reachability detection for SIP trunk destinations, rather than per-call reachability determination. **Early Offer support for voice and video calls Mandatory (insert MTP if needed)** and **Best Effort Early Offer** eliminate the use of MTPs to create an Early Offer for voice, video, and encrypted calls over SIP trunks. With **Best Effort Early Offer**, Unified CM sends only SIP Early Offer if the media characteristics of the calling device can be determined (for example, a call from a SIP-based IP phone over a Best Effort Early Offer trunk). If the media characteristics of the calling device cannot be determined (for example, for an inbound SIP Delayed Offer call forwarded over a Best Effort SIP Early Offer trunk), SIP Delayed Offer is sent instead.

SIP trunk normalization and transparency using Lua scripts improve native Unified CM interoperability with third-party unified communications systems. Normalization allows inbound and outbound SIP messages and SDP information to be modified on a per-SIP-trunk basis. Transparency allows Unified CM to pass SIP headers, parameters, and content bodies from one SIP trunk call leg to another, even if Unified CM does not understand or support the parts of the message that are being passed through.

These features are discussed in detail later in this section.

For the complete list of new enhancements for SIP trunks, refer to the Cisco Unified Communications Manager product release notes available at

https://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_release_notes_list.html

Session Initiation Protocol (SIP) Operation

This section explains how Unified CM SIP trunks operate and describes several key SIP trunk features that should be taken into account when designing and deploying Unified CM SIP trunks.

SIP Offer/Answer Model

Cisco Unified CM uses the SIP Offer/Answer model for establishing SIP sessions, as defined in RFC 3264. In this context, an Offer is contained in the Session Description Protocol (SDP) fields sent in the body of a SIP message. The Offer typically defines the media characteristics supported by the device (media streams, codecs, directional media attributes, IP address, and ports to use). The device receiving the Offer sends an Answer in the SDP fields of its SIP response, with its corresponding matching media streams, codec, directional media attributes, and the IP address and port number on which it wants to receive the media streams. Once the Offer and Answer have been exchanged, two-way media can be established between the calling and called endpoints. Unified CM uses this Offer/Answer model to establish SIP sessions as defined in the key SIP standard, RFC 3261.

RFC 3261 defines two ways that SDP messages can be sent in the Offer and Answer. These two methods are commonly known as Delayed Offer and Early Offer, and support for both methods by User Agent Client/Servers is required by specification RFC 3261. In the simplest terms, an initial SIP Invite sent with SDP in the message body defines an Early Offer, whereas an initial SIP Invite without SDP in the message body defines a Delayed Offer.

Delayed Offer and Early Offer are the two options available to all standards-based SIP switches for media capabilities exchange. Most vendors have a preference for either Delayed Offer or Early Offer, each of which has its own set of benefits and limitations.

Unified CM SIP trunks support both SIP Delayed Offer and SIP Early Offer. By default, SIP trunks are configured as Delayed Offer and support voice, video and encrypted calls. For Early Offer calls, there are three possible trunk configuration options:

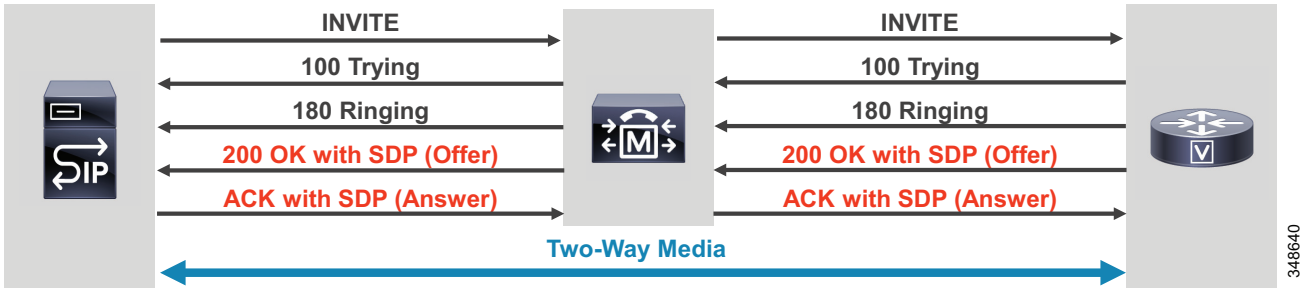
- **MTP Required** option selected on the SIP trunk — An MTP is inserted for every call.
- **Early Offer support for voice and video calls Mandatory (insert MTP if needed)** — A SIP Profile option, where Unified CM inserts a media termination point (MTP) if the media characteristics of the calling device cannot be determined (for example, for an inbound Delayed Offer call forwarded over an Early Offer SIP trunk).
- **Early Offer support for voice and video calls Best Effort (no MTP inserted)** — A SIP Profile option, where an Early Offer is sent only if the media characteristics of the calling device can be determined. If the media characteristics cannot be determined, a Delayed Offer is sent.

Unified CM Early Offer trunk configuration for Delayed Offer, Early Offer, and Best Effort Early Offer is discussed in the section on Unified CM SIP Trunks – Delayed Offer, Early Offer, and Best Effort Early Offer, page 6-18.

SIP Delayed Offer

In a Delayed Offer, the session initiator (calling device) does not send its capabilities in the initial Invite but waits for the called device to send its capabilities first (for example, the list of codecs supported by the called device, thus allowing the calling device to choose the codec to be used for the session). Figure 6-2 shows an example of a SIP Delayed Offer.

Figure 6-2 SIP Delayed Offer

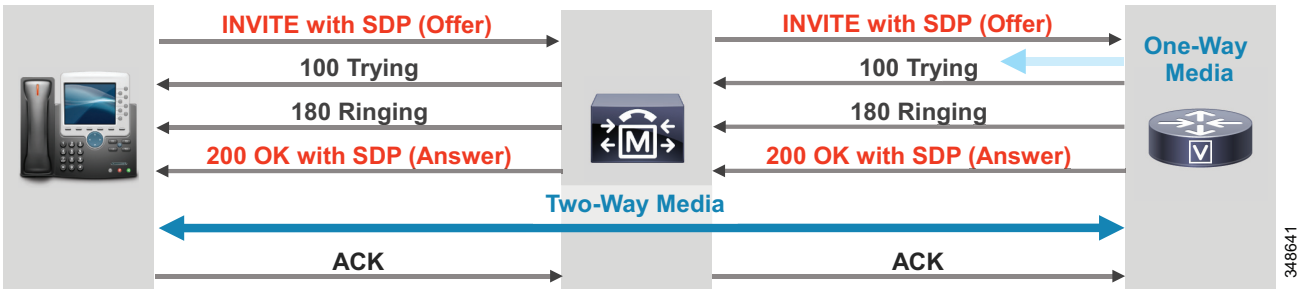


SIP Early Offer

In an Early Offer, the session initiator (calling device) sends its capabilities (for example, codecs supported, IP address, and UDP port number for RTP) in the SDP body contained in the initial Invite (thus allowing the called device to choose the codec for the session). Although both Early Offer and Delayed Offer are mandatory parts of the SIP standard, Early Offer is often preferred by third-party unified communications vendors, and it is almost always used by IP PSTN service providers. Service providers use a feature of Early Offer that allows one-way media to be established to the calling device on receipt of the SDP Offer in the initial INVITE. This one-way media feature is used to play announcements to the caller (for example, Unknown Number) before call charges commence. (Call charges typically commence after two-way media is established and the final acknowledgment (ACK) for the transaction is received.)

Note SIP-based Cisco Unified IP Phones send Early Offer. (See Figure 6-3.)

Figure 6-3 SIP Early Offer



Provisional Reliable Acknowledgement (PRACK)

SIP defines two types of responses to SIP Requests: Final Responses and Provisional Responses.

Final Responses (for example, 2XX, 3XX, and 4XX Responses) convey the result of a processed Request (such as an INVITE) and are sent reliably (which means they are acknowledged).

Provisional Responses (all 1XX Responses) provide information on the progress of the request, but they are not sent reliably, so the sender of a provisional response does not know that it has been received. For this reason SDP information is not sent in unreliable 1XX responses.

Provisional Reliable Acknowledgment (PRACK) is an extension to the SIP protocol that allows 1XX responses to be sent reliably. PRACK is useful because it provides reliability of 1XX responses for interoperability scenarios with the PSTN, and it can also be used to reduce the number of SIP messages that need to be exchanged before setting up two-way media. (See Figure 6-4 and Figure 6-5.)

PRACK can be used over SIP trunks using Early Offer or Delayed Offer, and it is often called *Early Media*. PRACK is supported by the majority of Cisco Collaboration products and is a generally recommended feature.

Figure 6-4 SIP Early Offer with Early Media (PRACK)

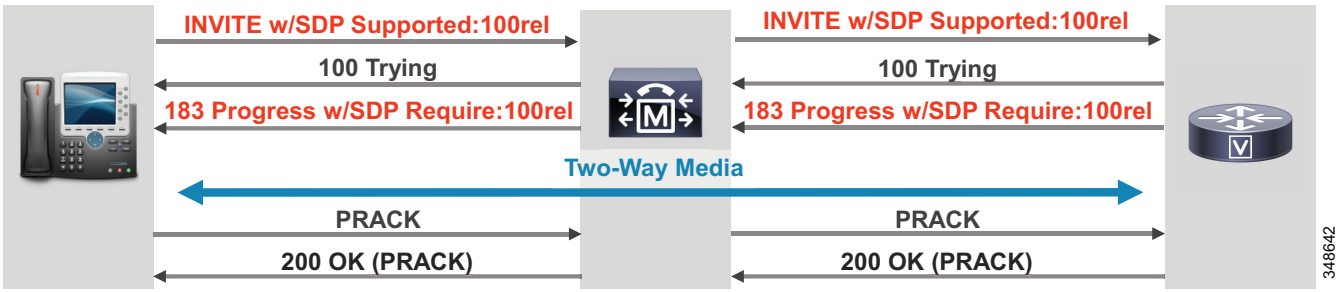
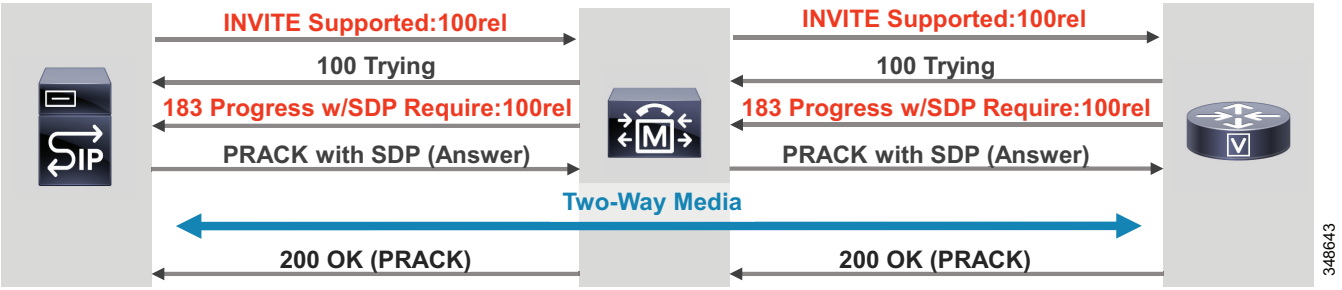


Figure 6-5 SIP Delayed Offer with Early Media (PRACK)



Note 100 Trying Responses indicate that Unified CM has received the INVITE. 180 Ringing and 183 Session in Progress Responses indicate that the user is being alerted of the call and are used to send information about the called user in SIP header messages and, if PRACK is used, in the SDP content in SIP message bodies.

Session Description Protocol (SDP) and Media Negotiation

SDP is the companion protocol of SIP. Defined in RFC 4566, SDP is used to describe media characteristics and to negotiate the media type, format, and associated parameters for a multimedia session between endpoints. These media characteristics are described by a series of one-line fields in a SDP message.

Session Description Protocol (SDP) and Voice Calls

The example in Table 6-2, Table 6-3, and Figure 6-6 illustrates an SDP Offer and Answer for a voice call.

Table 6-2 Voice Call – SDP Offer

SDP Message Field	Description
v=0	SDP version (currently version 0)
o=CiscoCCM-SIP 2000 1 IN IP4 10.10.199.250	Origin (contains Unified CM IP address)
s=SIP Call	Session name
c=IN IP4 10.10.199.130	Connection data (endpoint IP address)
t=0 0	Timing (0 0 = permanent session)
m=audio 16444 RTP/AVP 0 8 18 101	Media description – UDP port, RTP payload type for offered codecs (in preference order), and DTMF
a=rtpmap:0 PCMU/8000	G.711 mu-law codec
a=ptime:20	Packetization (sampling) interval (ms)
a=rtpmap:8 PCMA/8000	G.711 a-law codec
a=ptime:20	Packetization (sampling) interval (ms)

Table 6-2 Voice Call – SDP Offer (continued)

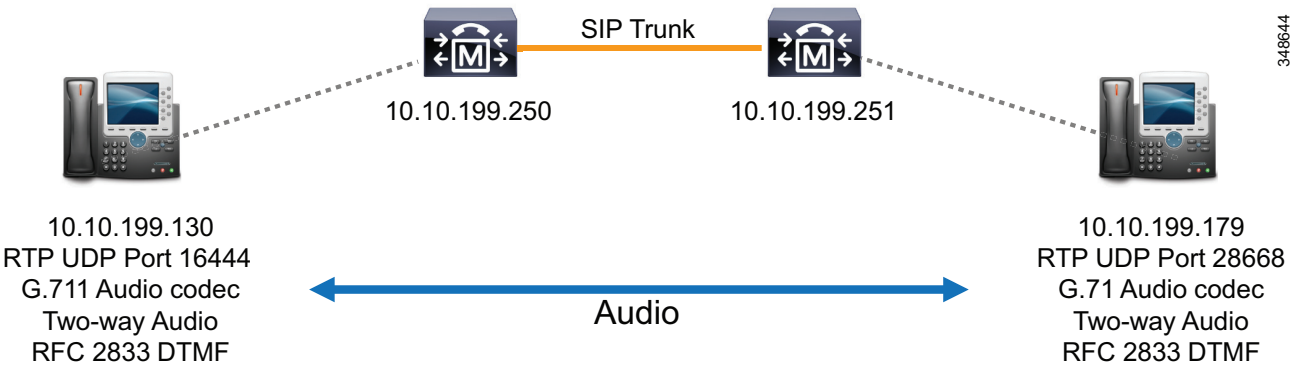
SDP Message Field	Description
a=rtpmap:18 G729/8000	G.729 codec
a=ptime:20	Packetization (sampling) interval (ms)
a=sendrecv	Media direction
a=rtpmap:101 telephone-event/8000	RFC 2833 in-band DTMF
a=fmtp:101 0-15	DTMF characters supported

The corresponding SDP Answer describes the media characteristics of the endpoint that receives the Offer and the voice codec selected by the endpoint for two-way voice media (see Table 6-3).

Table 6-3 Voice Call – SDP Answer

SDP Message Field	Description
v=0	SDP version (currently version 0)
o=CiscoCCM-SIP 2000 1 IN IP4 10.10.199.251	Origin (contains Unified CM IP address)
s=SIP Call	Session name
c=IN IP4 10.10.199.179	Connection data (endpoint IP address)
t=0 0	Timing (0 0 = permanent session)
m=audio 28668 RTP/AVP 0 101	Media description – UDP port, RTP payload type for the selected codec, and DTMF
a=rtpmap: 0 PCMU/8000	G.711 mu-law codec
a=ptime:20	Packetization (sampling) interval (ms)
a=sendrecv	Media direction
a=rtpmap:101 telephone-event/8000	RFC 2833 in-band DTMF
a=fmtp:101 0-15	DTMF characters supported

Figure 6-6 Negotiated Voice Call



Session Description Protocol (SDP) and Video Calls

For voice calls, symmetric media flows with a common voice codec are negotiated by the endpoints. For video media flows, it is commonly desirable for the send and receive media capabilities to be asymmetric. The requirement for asymmetry stems from a number of use cases such as broadband services where the upload and download speeds are different (often by an order of magnitude). In addition, video encoding is more CPU intensive than decoding video, and video endpoints can typically decode at higher resolution than they can encode. Because of these requirements, the video codec capabilities sent in an SDP Offer and Answer should be considered as the receive capabilities of the respective endpoints and are commonly asymmetric.

Table 6-4 shows the SDP Offer for a voice and video call.

Table 6-4 Voice and Video Call – SDP Offer

SDP Message Field	Description
v=0	SDP version (currently version 0)
o=CiscoCCM-SIP 161095 1 IN IP4 10.10.199.250	Origin (contains Unified CM IP address)
s=SIP Call	Session name
t=0 0	Timing (0 0 = permanent session)
m=audio 16444 RTP/AVP 0 8 18 101	Audio media – Port number and audio codecs listed by payload type in preference order and DTMF payload type
c=IN IP4 10.10.199.130	Connection data (endpoint IP address)
….	Attributes of multiple audio codecs and DTMF
m=video 16446 RTP/AVP 98 99	Media description – UDP port and RTP payload type for offered video codecs (in preference order)
c=IN IP4 10.10.199.130	Endpoint IP address
a=rtpmap:98 H264/90000	H.264 video codec
a=fmtp:98 profile-level-id=428016;packetization-mode=1;max-mbps=245000;max-fs=9000;max-cpb=200;max-br=5000;max-rcmd-nalu-size=3456000;max-smbps=245000;max-fps=6000	H.264 codec media attributes
a=rtpmap:99 H263-1998/90000	H.263 video codec
a=fmtp:99 QCIF=1;CIF=1;CIF4=1;CUSTOM=352,240,1	H.263 codec media attributes
a=rtcp-fb:* nack pli	RTCP for packet loss indication
a=rtcp-fb:* ccm tmmbr	RTCP for video rate adaptation

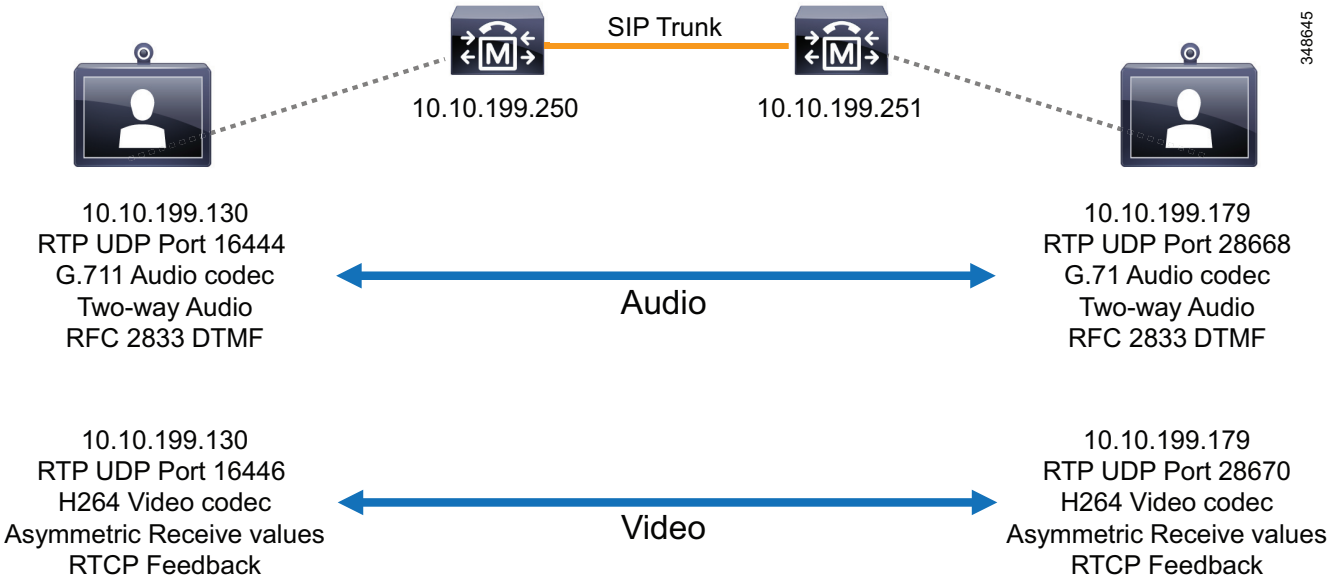
Notice that the H.264 and H.263 codecs offered in this SDP message contain a range of additional parameters that describe the receive capabilities of the endpoint. As shown in Table 6-5 for the negotiated H.264 codec in the SDP Answer, these parameters do not need to be symmetrical.

Table 6-5 Voice and Video Call – SDP Answer

SDP Message Field	Description
v=0	SDP version (currently version 0)
o=CiscoCCM-SIP 112480 1 IN IP4 10.10.199.251	Origin (contains Unified CM IP address)
s=SIP Call	Session name
t=0 0	Timing (0 0 = permanent session)
m=audio 28668 RTP/AVP 0 101	Audio media – Port number, selected audio codec, and DTMF payload type
c=IN IP4 10.10.199.179	Connection data (endpoint IP address)
….	Attributes of selected G.711 audio codec and DTMF
m=video 28670 RTP/AVP 98	H.264 codec selected for video
c=IN IP4 10.10.199.179	Endpoint IP address
a=rtpmap:98 H264/90000	H.264 codec details
a=fmtp:98 profile-level-id=428016;packetization-mode=1;max-mbps=108000;max-fs=3600;max-cpb=200;max-br=5000;max-rcmd-nalu-size=1382400;max-smbps=108000;max-fps=6000	Media attributes of the selected H.264 codec. Profile-level-id and packetization mode must be symmetric for the negotiated call; other attributes need not be symmetric and represent the receive capabilities of the endpoint.
a=rtcp-fb:* nack pli	RTCP for packet loss indication
a=rtcp-fb:* ccm tmmbr	RTCP for video rate adaptation

The Profile Level ID and Packetization Mode must be symmetrical for the negotiated video call. The Profile Level ID describes a minimum subset of H.264 features, resolution, frame rate, and bit rate supported by the endpoint. The Packetization Mode describes how video samples can be encapsulated and sent in RTP packets. The media attributes, which follow the Profile Level ID and Packetization Mode, need not be symmetrical and indeed are not all symmetrical for the negotiated video call shown in Table 6-5 and Figure 6-7.

Figure 6-7 Negotiated Voice and Video Call



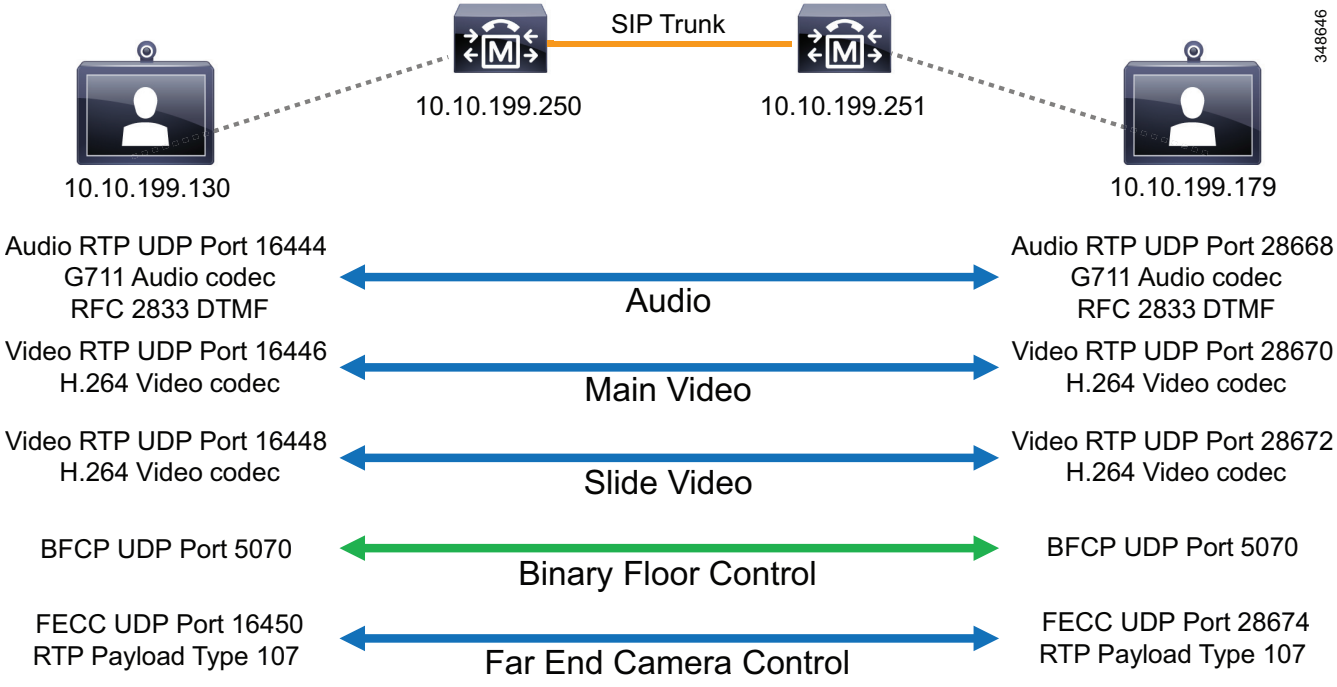
Video Desktop Sharing and Binary Floor Control Protocol (BFCP)

For video desktop and presentation sharing, the endpoints negotiate an additional RTP video channel to transmit the shared content (presentation slides, for example) and a UDP channel for BFCP, which manages shared access to resources within the video or conference call. (See Figure 6-8.) BFCP is described in RFC 4582 and RFC 4583.

Far End Camera Control (FECC)

Far End Camera Control allows a user to select a video source and to control camera actions such as pan, tilt, zoom and focus. Endpoints using FECC negotiate an additional RTP channel for camera control. (See Figure 6-8.) FECC is described in H.281, H.224, and RFC 4573.

Figure 6-8 Audio and Video Call with Presentation Sharing and Far End Camera Control



Unified CM SIP Trunk Features and Operation

This section explains how Unified CM SIP trunks operate and describes several key SIP trunk features that should be taken into account when designing and deploying Unified CM SIP trunks.

Run on All Unified CM Nodes

Cisco Unified CM provides a configuration option for allowing SIP trunk calls to be made or received on any call processing subscriber node in the cluster.

SIP Trunks – Run on All Nodes and the Route Local Rule

When the **Run on all Active Unified CM Nodes** option is checked on a SIP trunk, Unified CM creates an instance of the SIP trunk daemon on every call processing subscriber within the cluster, thus allowing a SIP trunk call to be made or received on any call processing subscriber. (Prior to this feature, up to three nodes could be selected per trunk by using Unified CM Groups.) With **Run on all Active Unified CM Nodes** enabled, outbound SIP trunk calls originate from the same node on which the inbound call (for example, from a phone or trunk) is received (based on the Route Local rule). The **Run on all Active Unified CM Nodes** feature overrides the trunk's Unified CM Group configuration.

For SIP trunks, the Route Local rule operates as follows:

> For outbound SIP trunk calls, when a call from a registered phone or inbound trunk arrives at a Unified CM node, Unified CM checks to see if an instance of the selected outbound trunk exists on the same node where the inbound call arrived. If so, Unified CM uses this node to establish the outbound trunk call.

Enabling **Run on all Active Unified CM Nodes** on SIP trunks is highly recommended because this feature allows outbound calls to originate from, and be received on, any call processing node within the cluster. **Run on all Active Unified CM Nodes** can also eliminate calls from being set up between call processing nodes within the same cluster before being established over the outbound SIP trunk.

As with all Unified CM SIP trunks, the SIP daemons associated with the trunk will accept inbound calls only from end systems with IP addresses that are defined in the trunk's destination address fields. When multiple SIP trunks to the same destination(s) are using the same call processing nodes, a unique incoming and destination port number must be defined per trunk to allow each trunk to be identified uniquely.

Route Lists – Run on All Nodes and the Route Local Rule

Although this is not specifically a SIP trunk feature, running route lists on all nodes provides benefits for trunks in route lists and route groups. Running route lists on all nodes improves outbound call distribution by using the Route Local rule to avoid unnecessary intra-cluster call setup traffic.

For route lists, the Route Local rule operates as follows:

> For outbound calls that use route lists (and associated route groups and trunks), when a call from a registered phone or inbound trunk arrives at the node with the route list instance, Unified CM checks to see if an instance of the selected outbound trunk exists on the same node as the route list. If so, Unified CM uses this node to establish the outbound trunk call.

If both the route list and the trunk have **Run on all Active Unified CM Nodes** enabled, outbound call distribution will be determined by the node on which the inbound call arrives. If the selected outbound trunk uses Unified CM Groups instead of running on all nodes, Unified CM applies the Route Local rule if an instance of the selected outbound trunk exists on the same node on which the inbound call arrived. If an instance of the trunk does not exist on this node, then Unified CM forwards the call (within the cluster) to a node where the trunk is active.

If the route list does not have **Run on all Active Unified CM Nodes** enabled, an instance of the route list will be active on one node within the cluster (the primary node of the route list's Unified CM Group). If the selected outbound trunk is also active on the primary node of the route list's Unified CM Group, the Route Local rule will apply, resulting in sub-optimal outbound call distribution because all outbound trunk calls will originate from this node.

Cisco strongly recommends enabling **Run on all Active Unified CM Nodes** on all route lists and SIP trunks.

Up to 16 SIP Trunk Destination IP Addresses

SIP trunks can be configured with up to 16 destination IP addresses, 16 fully qualified domain names, or a single DNS SRV entry. Support for additional destination IP addresses reduces the need to create multiple trunks associated with route lists and route groups for call distribution between two Unified Communications systems, thus simplifying Unified CM trunk design. This feature can be used in conjunction with the **Run on all Active Unified CM Nodes** feature. (See Figure 6-9 and Figure 6-10.) Bear in mind, however, that the SIP daemons associated with a Unified CM SIP trunk will accept inbound calls only from end systems with IP addresses that are defined in the trunk's destination address

fields. Use a single SIP trunk with one or more destination addresses to connect a Unified CM cluster to one other unified communications system. If trunk fail-over is required, create an additional trunk to the fail-over unified communications system and use route lists and route groups to order trunk selection. Unified CM randomly distributes outbound calls over the configured SIP trunk destination addresses.

Figure 6-9 SIP Trunks with Run on All Unified CM Nodes and Multiple Destination Addresses

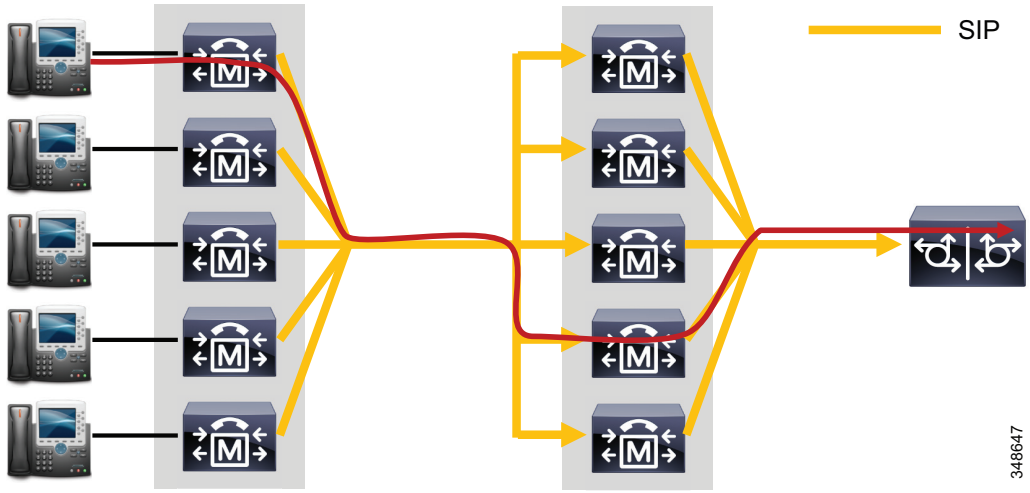
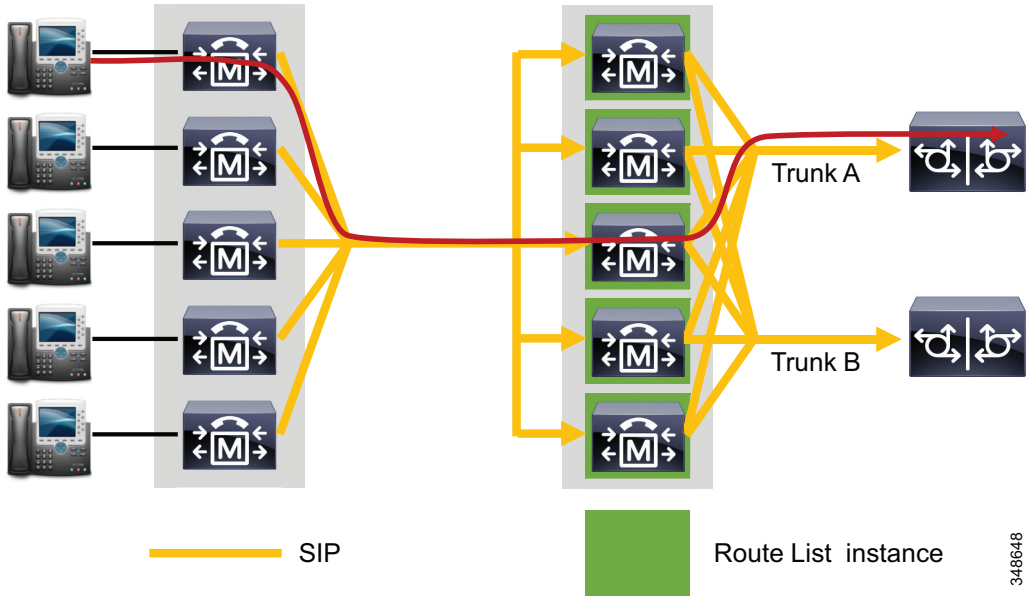


Figure 6-10 SIP Trunks and Route Lists with Run on All Unified CM Nodes Enabled



SIP Trunks Using DNS

Using a DNS SRV entry as the destination of a SIP trunk might be preferable to defining multiple destination IP addresses in certain situations such as the following:

- SRV host prioritization is required
- SRV host weighting is required
- More than 16 destination IP addresses are required
- DNS SRV resolution is a requirement of the destination Unified Communications system

Note If the configuration option **Destination Address is an SRV** is selected, only a single SRV entry can be added as the trunk destination. (For example, Destination Address = cluster1.cisco.com. Port = 0.)

Figure 6-11 shows the call flow for a SIP trunk using DNS SRV to resolve the addresses to a destination Unified CM cluster. However, this destination could also be a third-party unified communications system.

Figure 6-11 Call Flow for Intercluster SIP Trunk Using DNS SRV

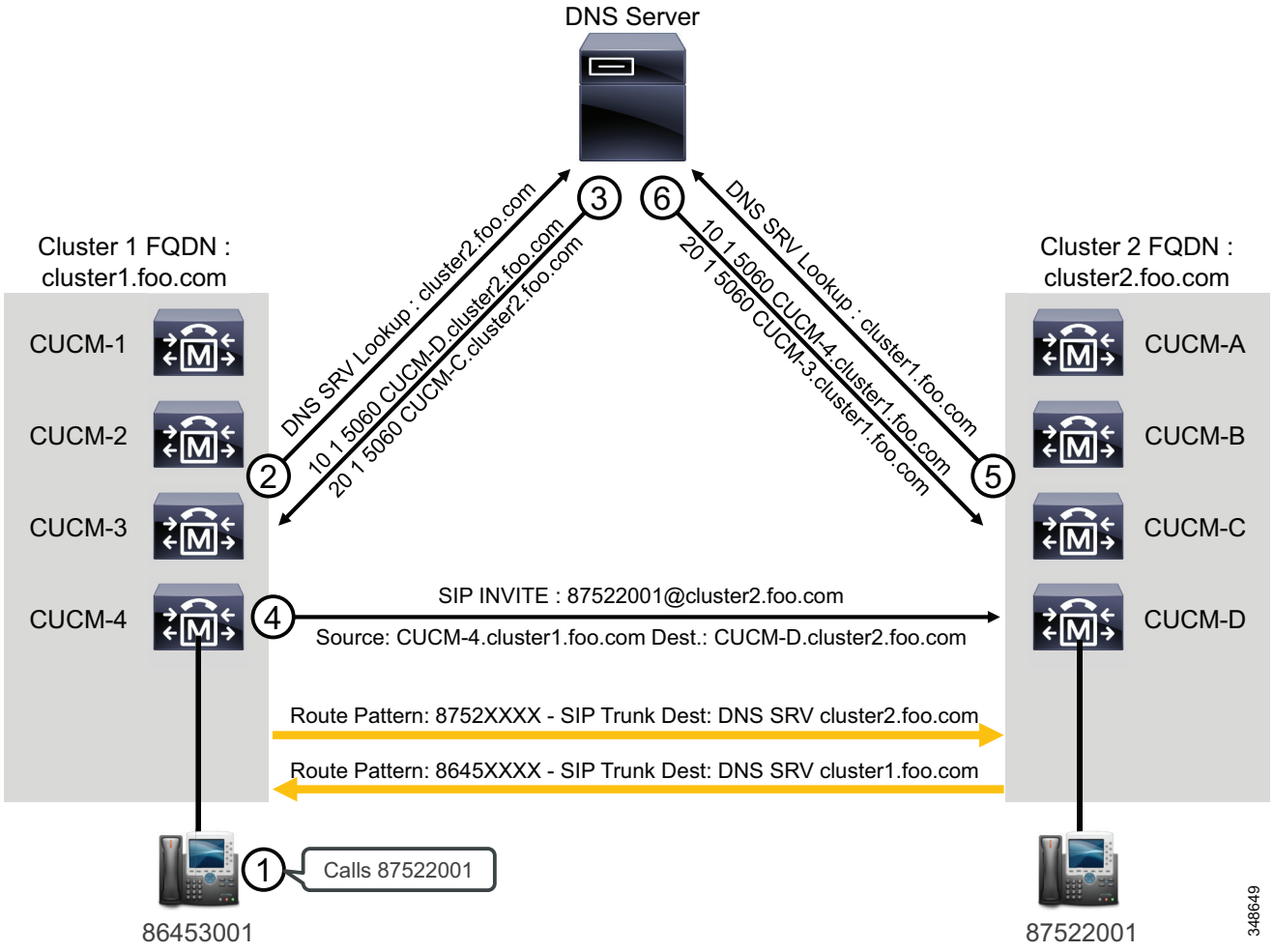


Figure 6-11 illustrates the following steps in the call flow:

1. The IP phone in Cluster 1 calls 87522001.
2. The call matches a route pattern of 8752XXXX that is pointing to the SIP trunk with DNS SRV of cluster2.foo.com. CUCM-4 in Cluster 1 is the node handling this call because the phone and the SIP trunk are both registered to it. CUCM-4 sends a DNS SRV lookup for cluster2.foo.com.
3. The DNS server replies with two records: CUCM-D.cluster2.foo.com and CUCM-C.cluster2.foo.com. Because CUCM-D.cluster2.foo.com has a higher priority, the call is attempted to this Unified CM. Before sending the SIP Invite, another DNS lookup is done for CUCM-D.cluster2.foo.com.
4. CUCM-4 sends a SIP Invite to 87522001@cluster2.foo.com, with destination address set to the IP address of CUCM-D.
5. Unified CM interprets this call as a local call because the host portion of the uniform resource identifier (URI) matches the Cluster FQDN enterprise parameter. Cluster 2 does not have any SIP trunk configured with a destination of CUCM-4, so it does a DNS SRV lookup for all domains configured under the SIP trunks with DNS SRV. In this case, the example shows a single trunk with a DNS SRV destination of cluster1.foo.com.
6. The DNS server returns two entries, and one of them matches the source IP address of the Invite. The cluster accepts the call and extends it to extension 87522001.

Note The DNS A Look-up is not shown in this call flow.

SIP OPTIONS Ping

The SIP OPTIONS Ping feature can be enabled on the SIP Profile associated with a SIP trunk to dynamically track the state of the trunk's destination(s). When this feature is enabled, each node running the trunk's SIP daemon will periodically send an OPTIONS Request to each of the trunk's destination IP addresses to determine its reachability and will send calls only to reachable nodes. A destination address is considered to be "out of service" if it fails to respond to an OPTIONS Request, if it sends a Service Unavailable (503) response or Request Timeout (408) response, or if a TCP connection cannot be established. The overall trunk state is considered to be "in service" when at least one node receives a response (other than a 408 or 503) from a least one destination address. SIP trunk nodes can send OPTIONS Requests to the trunk's configured destination IP addresses or to the resolved IP addresses of the trunk's DNS SRV entry. Enabling SIP OPTIONS Ping is recommended for all SIP trunks because it allows Unified CM to track trunk state dynamically rather than determining trunk destination state on a per-node, per-call, and time-out basis.

Unified CM SIP Trunks – Delayed Offer, Early Offer, and Best Effort Early Offer

This section provides guidance on the use of Delayed Offer, Early Offer, and Best Effort Early Offer with Unified CM SIP trunks.

Unified CM SIP Delayed Offer

The default configuration for Unified CM SIP trunks is to use Delayed Offer (SIP INVITE sent without SDP content). Using this default configuration, all outbound calls over the SIP trunk send SIP Delayed Offer. Media termination points (MTPs) are not used in the outbound INVITE or to generate SDP content

in the Answer sent in response to a received Offer. However, MTPs may be used to address DTMF transport mismatches. Use this default configuration if you want all calls sent over the SIP trunk to send Delayed Offer. Voice, video and encrypted calls are supported.

Unified CM SIP Early Offer

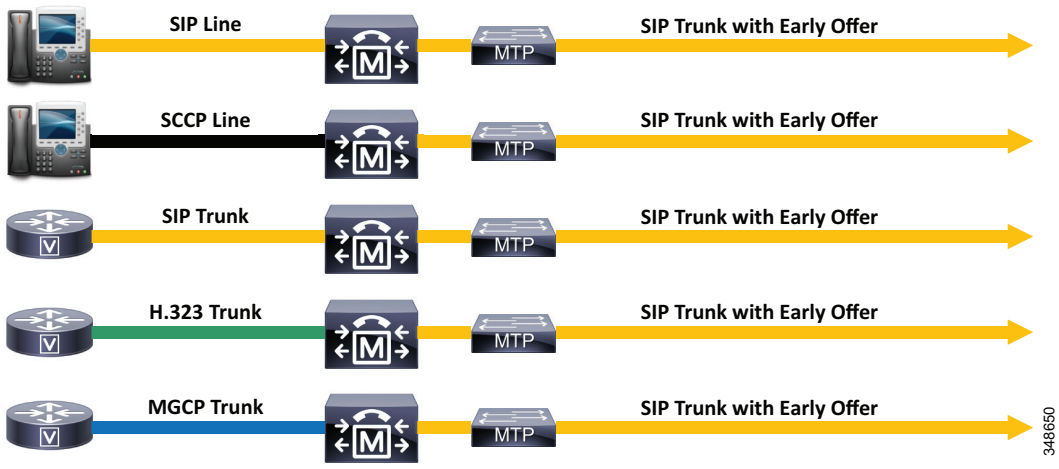
Two configurable options are available to enable Early Offer for all outbound calls over Unified CM SIP trunks:

- Media Termination Point Required
- Early Offer support for voice and video calls Mandatory (insert MTP if needed)

Early Offer Using Media Termination Point Required

Enabling the **Media Termination Point Required** option on the SIP trunk assigns an MTP from the trunk's media resources group (MRG) to every inbound and outbound call. (See Figure 6-12.) This statically assigned MTP supports either the G.711 or G.729 codec only, thus limiting media to voice calls only, using the selected codec type. Enabling Early Offer using **Media Termination Point Required** has been superseded by **Early Offer support for voice and video calls Mandatory (insert MTP if needed)** and **Early Offer support for voice and video calls Best Effort (no MTP inserted)**. Early Offer using **Media Termination Point Required** can be useful in cases where voice media for inbound and outbound calls needs to be anchored to a single IP address (that of the MTP).

Figure 6-12 SIP Early Offer with Media Termination Point Required



Early Offer Using Early Offer support for voice and video calls Mandatory (insert MTP if needed)

Enabling **Early Offer support for voice and video calls Mandatory (insert MTP if needed)** on the SIP Profile associated with the SIP trunk inserts an MTP only if the calling device cannot provide Unified CM with the media characteristics required to create the Early Offer. In general, **Early Offer support for voice and video calls Mandatory (insert MTP if needed)** is recommended over **Media Termination Point Required** because this configuration option reduces MTP usage and can support voice, video, and encrypted calls. (see Figure 6-13).

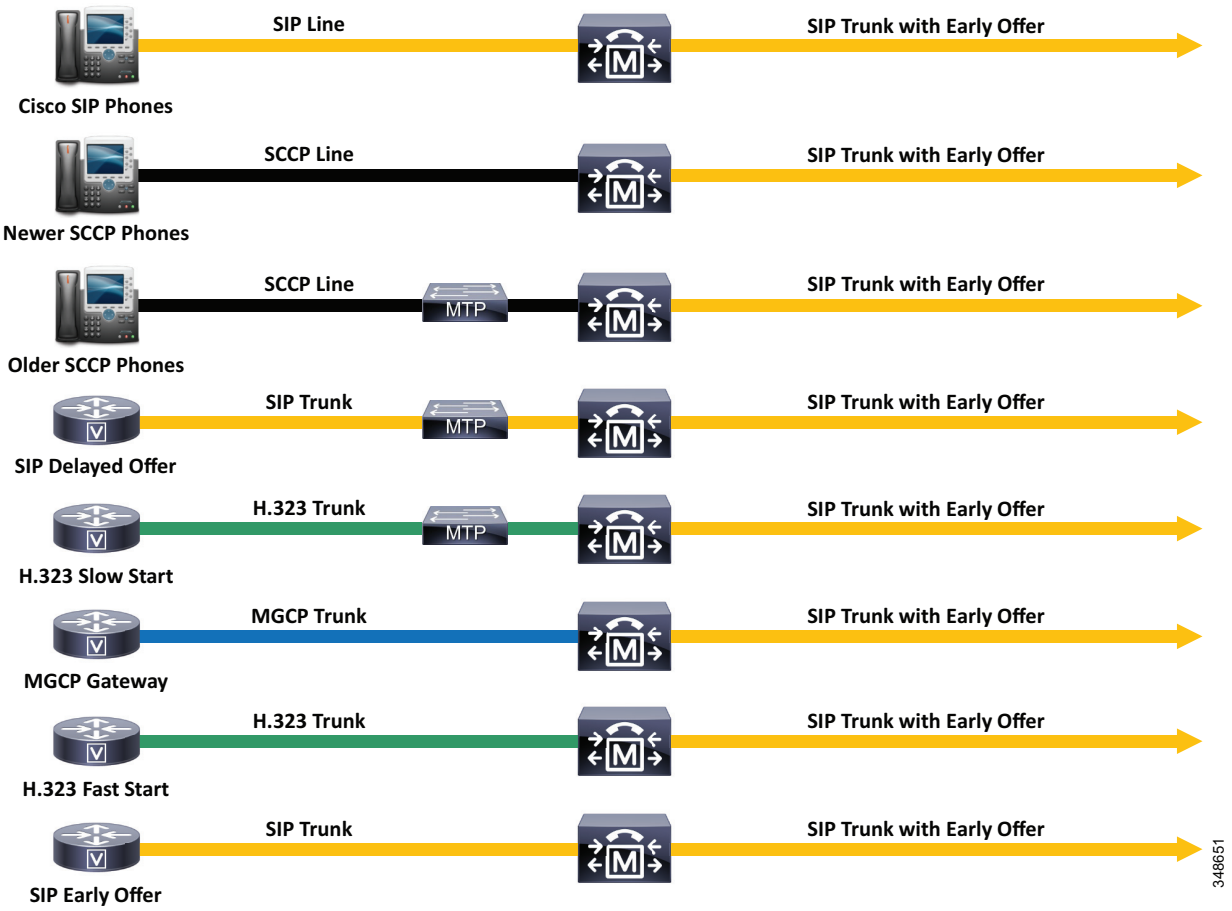
For outbound calls over a SIP trunk configured as Early Offer support for voice and video calls Mandatory (insert MTP if needed), Unified CM inserts an MTP to create an SDP Offer in the following cases only:

- An inbound call to Unified CM is received over a Delayed Offer SIP trunk
- An inbound call to Unified CM is received over an H.323 Slow Start trunk
- An inbound call is received from an older SCCP-based IP phone registered to Unified CM

As a general rule, Early Offer calls of this type that use MTPs support voice only, but they are not limited to a single voice codec (as they are with Early Offer using MTP Required). These calls support only audio in the initial call setup but can be escalated mid-call to support video and SRTP if the call media is renegotiated (for example, after hold or resume).

Note MTP resources are not required for incoming INVITE messages, whether or not they contain an initial Offer SDP.

Figure 6-13 Early Offer support for voice and video calls - Mandatory (insert MTP if needed)



Unified CM does not need to insert an MTP to create an outbound Early Offer call over a SIP trunk if the inbound call to Unified CM is received by any of the following means:

- On a SIP trunk using Early Offer
- On an H.323 trunk using Fast Start
- On an MGCP trunk
- From a SIP-based IP phone registered to Unified CM
- From newer SCCP-based Cisco Unified IP Phone models registered to Unified CM

For the above devices, Unified CM uses the media capabilities of the endpoint and applies the codec filtering rules based on the region-pair of the calling device and outgoing SIP trunk to create the offer SDP for the outbound SIP trunk call. In most cases, the offer SDP will have the IP address and port number of the endpoint initiating the call. This is assuming that Unified CM does not have to insert an MTP for other reasons such as a DTMF mismatch, TRP requirements, or a transcoder requirement when there is no common codec between the regions of the calling device and the SIP trunk.

When **Early Offer support for voice and video calls Mandatory (insert MTP if needed)** is configured on a trunk's SIP Profile, calls from older SCCP-based phones, SIP Delayed Offer trunks, and H.323 Slow Start trunks will cause Unified CM to allocate an MTP. The MTP is used to generate an offer SDP with a valid media port and IP address. The MTP will be allocated from the media resources associated with the calling device rather than from the outbound SIP trunk's media resources. (This prevents the media path from being hair-pinned via the outbound SIP trunk's MTP). If the MTP cannot be allocated from the calling device's media resource group list (MRGL), then the MTP allocation is attempted from the SIP trunk's MRGL.

For calls from older SCCP phones registered to Unified CM, some of the media capabilities of the calling device (for example, supported voice codecs, video codecs, and encryption keys if supported) are available for media exchange through the Session Description Protocol (SDP). Unified CM will create a superset of the endpoint and MTP codec capabilities and apply the codec filtering based on the applicable region-pair settings. The outbound Offer SDP will use the MTP's IP address and port number and can support voice, video, and encrypted media. Note that a Cisco IOS-based MTP should be used and configured to support the pass-through codec.

Note Older SCCP-based IP phones such as the Cisco Unified IP Phone 7902, 7905, 7910, 7912, 7920, 7935, 7940, and 7960 require the use of an MTP when they make calls over a SIP trunk with the **Early Offer for voice and video Mandatory (insert MTP if needed)** feature enabled. If you have a significant number of these phone types deployed in a cluster, consider deploying Delayed Offer trunks instead of **Early Offer for voice and video Mandatory (insert MTP if needed)**. If **Early Offer for voice and video Mandatory (insert MTP if needed)** trunks are used, provision MTP resources in the cluster equivalent to the number of busy hour calls over those SIP trunks that use this Early Offer feature.

When Unified CM receives an inbound call on an H.323 Slow Start or SIP Delayed Offer trunk, the media capabilities of the calling device are not available when the call is initiated. In this case, Unified CM must insert an MTP and will use its IP address and UDP port number to advertise all supported audio codecs (after region pair filtering) in the Offer SDP of the initial INVITE sent over the outbound SIP trunk. When the Answer SDP is received on the SIP trunk, if it contains a codec that is supported by the calling endpoint, then no additional offer-answer transaction is needed. In case of codec mismatch, Unified CM can either insert a transcoder to address the mismatch or send a Re-INVITE or UPDATE to trigger media negotiation. Calls from H.323 Slow Start or SIP Delayed Offer trunks support only audio in the initial call setup, but they can be escalated mid-call to support video and SRTP if the call media is renegotiated (for example, after Hold or Resume).

Best Effort Early Offer [Early Offer support for voice and video calls Best Effort (no MTP inserted)]

Best Effort Early Offer can be enabled on the SIP Profile associated with the SIP trunk, and it is the recommended configuration for all Unified CM and Unified CM Session Management Edition (SME) trunks. **Best Effort Early Offer** trunks never use MTPs to create an Early Offer and, depending on the calling device, may initiate an outbound SIP trunk call using either Early Offer or Delayed Offer. **Best Effort Early Offer** SIP trunks support voice, video, and encrypted calls.

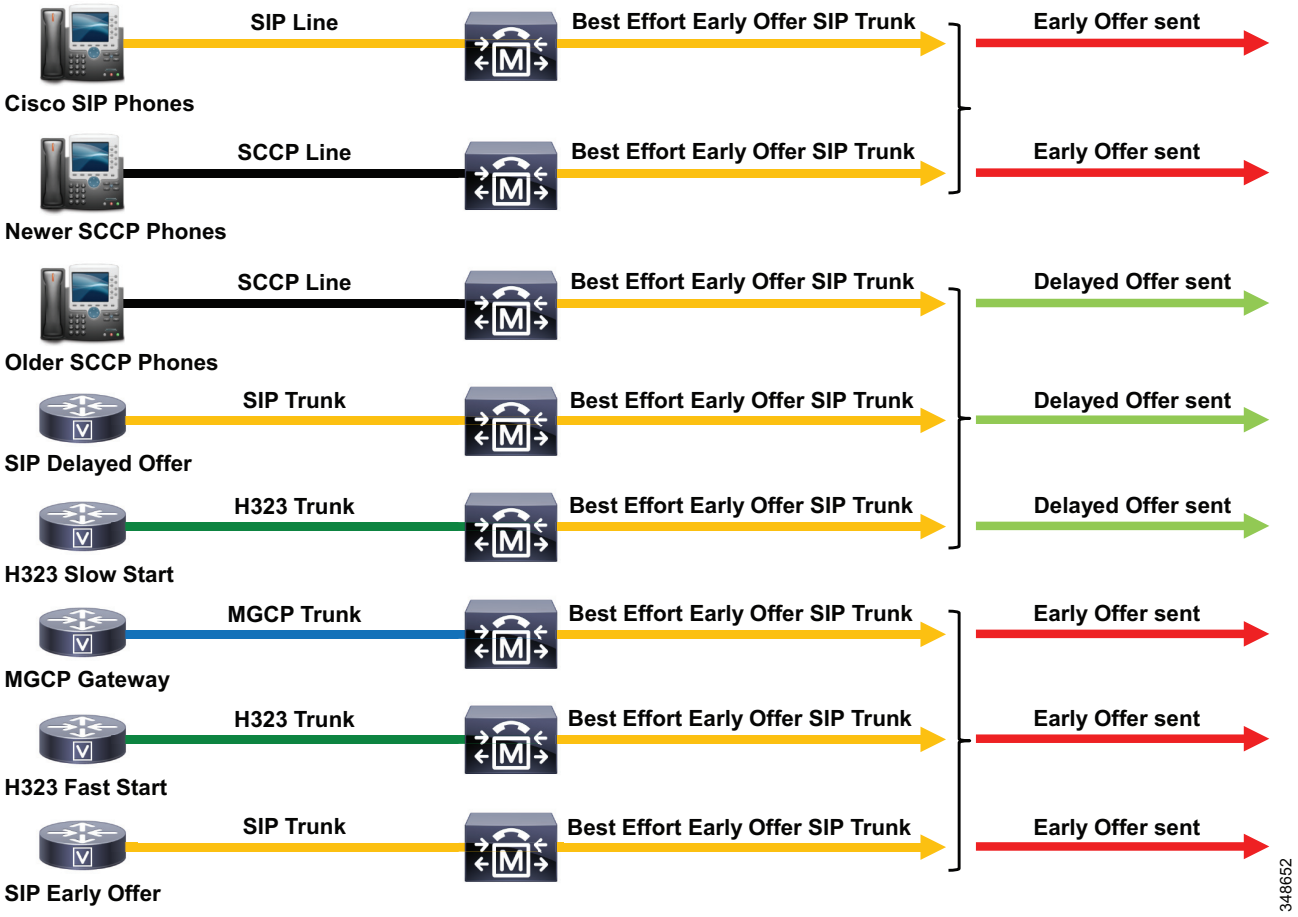
Best Effort Early Offer SIP trunks send outbound calls with an Early Offer (INVITE with SDP content) in the following situations:

- An inbound call to Unified CM or SME is received over a SIP trunk using Early Offer.
- An inbound call to Unified CM or SME is received over an H.323 trunk using Fast Start.
- An inbound call to Unified CM or SME is received over an MGCP trunk.
- A call is initiated from a SIP-based IP phone registered to Unified CM.
- A call is initiated from a newer model SCCP-based Cisco Unified IP Phone registered to Unified CM.

Best Effort Early Offer trunks send outbound calls with a Delayed Offer (INVITE without SDP content) in the following situations:

- An inbound call to Unified CM or SME is received over a Delayed Offer SIP trunk.
- An inbound call to Unified CM or SME is received over an H.323 Slow Start trunk.
- A call is initiated from an older model SCCP-based IP phone registered to Unified CM.

Figure 6-14 Best Effort Early Offer



Media resources such as MTPs for DTMF translation, trusted relay points (TRPs), and transcoders for codecs mismatches can still be associated with and used by a **Best Effort Early Offer** trunk. Note that with **Best Effort Early Offer**, MTPs are never used to create an Early Offer or to create an Answer in response to a received Offer.

Using **Best Effort Early Offer** for all SIP trunks in your enterprise simplifies Cisco Collaboration System network design and deployments, and it eliminates the need to use MTPs to generate an Offer. Note, however, that Cisco Collaboration call control systems, applications, and gateways may receive either an Early Offer or Delayed Offer call over a **Best Effort Early Offer** trunk, and they should be able to receive either. All Cisco Collaboration System applications support the receipt of either Early Offer or Delayed Offer calls.

In certain cases (for example, calls via a Cisco Unified Border Element Session Border Controller (SBC) to a service provider's IP PSTN), Early Offer must always be sent to the IP PSTN. In these situations, use Cisco Unified Border Element's Delayed Offer to Early Offer feature to convert a received Delayed Offer to Early Offer.

If your Cisco Collaboration System application must receive either Early Offer only or Delayed Offer only, you can use a Unified CM SIP trunk configured for Early Offer (using **Early Offer support for voice and video calls Mandatory (insert MTP if needed)** or **MTP Required**) or Delayed Offer, respectively, to connect to this application. With single Unified CM cluster deployments, these trunk

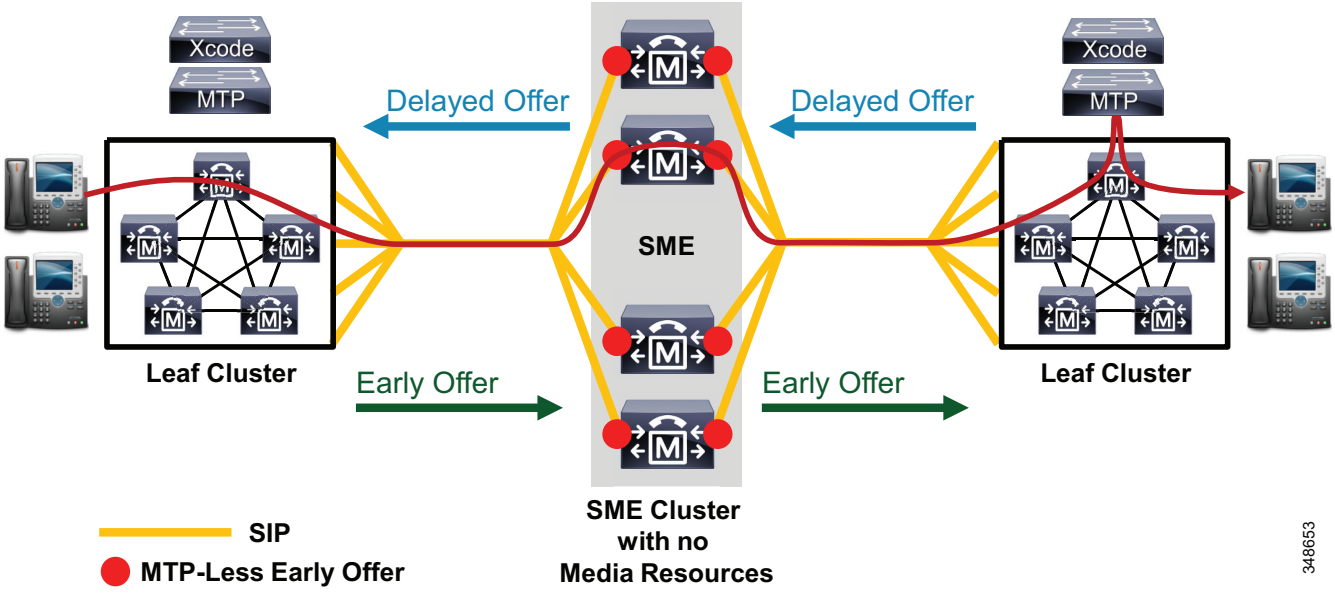
choices are straightforward. For multi-cluster deployments interconnected via Unified CM Session Management Edition, where a single SIP trunk can be shared to reach many end Cisco Collaboration Systems, **Best Effort Early Offer** is recommended for all SME trunks. For more information on the design considerations for **Best Effort Early Offer**, see Summary of SIP Trunk Recommendations for Multi-Cluster SME Deployments, page 6-47.

MTP-Less Early Offer, Best Effort Early Offer, and SME Media Transparency

MTP-Less Early Offer is a special SIP trunk configuration for Unified CM Session Manager Edition (SME) cluster versions that do not support the **Best Effort Early Offer** feature. **Best Effort Early Offer** provides the same functionality as **MTP-Less Early Offer**; but whereas **MTP-Less Early Offer** deployments require that no media resources are configured on the SME cluster, with **Best Effort Early Offer**, media resources can be configured if needed. SME deployments using only **Best Effort Early Offer** or **MTP-Less Early Offer** SIP trunks allow you to deploy an SME cluster that is media transparent (no media resources are required in the SME cluster) because all media negotiation takes place in the leaf Unified Communications systems, which insert media resources (MTPs, transcoders, and so forth) as required. (See Figure 6-15.)

MTP-Less Early Offer takes advantage the Unified CM SIP service parameter **Fail Call Over SIP Trunk if MTP Allocation Fails**. The default setting for this service parameter is **False**, thus allowing an inbound Delayed Offer call to proceed over the outbound SIP trunk (configured for Early Offer) as a Delayed Offer call if no MTP resources are available.

Figure 6-15 Using MTP-Less Early Offer for SME Media Transparency

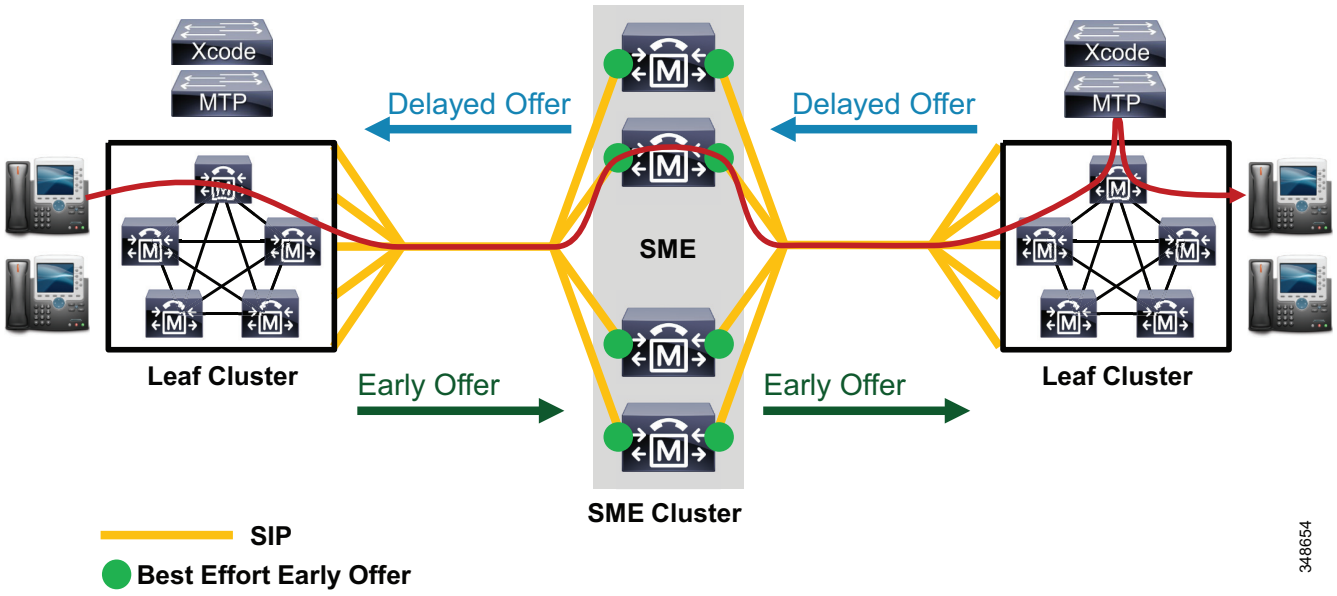


To configure a media-transparent SME cluster using MTP-less Early Offer:

- Use only SIP trunks on the SME cluster.
- Enable all trunks with **Early Offer support for voice and video calls Mandatory (insert MTP if needed)**.
- Disable the IPVMS service on all SME nodes. This disables Unified CM media termination points, conferencing, music on hold, and annunciator resources.

- Do not associate any Cisco IOS media resources with the SME cluster.
- Configure SIP trunk DTMF settings to **No Preference** (the default setting).
- Enable **Accept Audio Codec Preference in Received Offer** on all SME SIP trunks.

Figure 6-16 Using Best Effort Early Offer for SME Media Transparency



To configure a media-transparent SME cluster using **Best Effort Early Offer**:

- Use only SIP trunks on the SME cluster.
- Enable all trunks with **Best Effort Early Offer**.
- Configure SIP trunk DTMF settings to **No Preference** (the default setting).
- Enable **Accept Audio Codec Preference in Received Offer** on all SME SIP trunks.

Note Media resources can be deployed in an SME cluster where **Best Effort Early Offer** SIP trunks are configured, but these resources will be used only if one or more SIP trunks are configured as Delayed Offer or Early Offer. In these cases, calls to and from Early Offer or Delayed Offer trunks are not media transparent and can invoke media resources if a DTMF or codec mismatch is encountered.

Media Termination Points

MTPs are used by Unified CM for the following purposes:

- To deliver a SIP Early Offer over SIP trunks
- To address DTMF transport mismatches
- To act as an RSVP agent
- To act as a Trusted Relay Point (TRP)
- To provide conversion between IPv4 and IPv6 for RTP streams

MTPs are available in three forms:

- Software MTPs in Cisco IOS gateways — Available with any Cisco IOS T-train software release and scaling up to 5,000 sessions (calls) on the Cisco Aggregation Services Routers (ASR) 1000 Series with Route Processor RP2.
- Hardware MTPs in Cisco IOS gateways — Available with any Cisco IOS T-train software release, hardware MTPs use on-board DSP resources and scale calls according to the number of DSPs supported on the Cisco router platform.
- Cisco Unified CM software MTPs using the Cisco IP Voice Media Streaming Application on a Unified CM subscriber node

Cisco IOS MTPs are recommended over Unified CM MTPs because Cisco IOS MTPs provide additional scalability and greater functionality, such as support for additional codec types, multiple media streams, and the pass-through codec. (For details, see the section on Media Termination Point (MTP), page 7-7.)

The following example configuration is for a Cisco IOS software MTP:

```
!
sccp local Vlan5
sccp ccm 10.10.5.1 identifier 5 version 8.6.2
! Communications Manager IP address (10.10.5.1)
sccp
!
sccp ccm group 5
 bind interface Vlan5
 associate ccm 5 priority 1
 associate profile 5 register MTP000E83783C50
! MTP name (MTP000E83783C50) ... must match the Unified CM MTP name.
!
dspfarm profile 5 mtp
 description software MTP
 codec g711ulaw
 codec pass-through
 maximum sessions software 500
 associate application SCCP
```

DTMF Transport over SIP Trunks

There are several methods of transporting DTMF information between SIP endpoints. In general terms, these methods can be classified as out-of-band and in-band signaling. In-band DTMF transport methods send either raw or signaled DTMF tones within the RTP stream, and they need to be handled and interpreted by the endpoints that generate and/or receive them. Out-of-band signaling methods transport DTMF tones outside of the RTP path, either directly to and from the endpoints or through a call agent such as Cisco Unified CM, which interprets and/or forwards these tones as required.

Out-of-band (OOB) SIP DTMF signaling methods include Unsolicited Notify (UN), Information (INFO), and Key Press Mark-up Language (KPML). KPML (RFC 4730) is the OOB signaling method preferred by Cisco and is supported by Cisco Unified CM, Cisco IOS platforms (Release 12.4 and later), and most models of Cisco Unified IP Phones. INFO is not supported by Unified CM.

In-band DTMF transport methods send DTMF tones as either raw tones in the RTP media stream or as signaled tones in the RTP payload using RFC 2833. Among SIP product vendors, RFC 2833 has become the predominant method of sending and receiving DTMF tones and is supported by the majority of Cisco voice products.

Because in-band signaling methods send DTMF tones in the RTP media stream, the SIP endpoints in a session must either support the transport method used (for example, RFC 2833) or provide a method of intercepting this in-band signaling and converting it. If the two endpoints are using a back-to-back user agent (B2BUA) server for the call control (for example, Cisco Unified CM) and the endpoints negotiate different DTMF methods between each device and call control agent, then the call control agent determines how to handle the DTMF differences, either through MTP insertion or by OOB methods. With Unified CM, a DTMF transport mismatch (for example, in-band to out-of-band DTMF) is resolved by inserting a media termination point (MTP), which terminates the RTP stream with in-band DTMF signaling (RFC 2833), extracts the DTMF tones from the RTP stream, and forwards these tones out-of-band to Unified CM, where they are then forwarded to the endpoint supporting out-of-band signaling. For DTMF mismatches, the inserted MTP is always in the media path between the two endpoints. In-band DTMF packets are identified by their RTP Payload type, extracted by Unified CM, and converted to out-of-band DTMF, while RTP media packets pass transparently through the MTP.

In-band DTMF tones can also be transported as raw (audible) tones in the RTP media stream. This transport method is not widely supported by Cisco products and, in general, is not recommended as an end-to-end DTMF transport mechanism. In-band audio DTMF tones can generally be reproduced reliably when using high-bandwidth codecs such as G.711 a-law or mu-law, but they are not suitable for use with low-bandwidth codecs such as G.729. In cases where in-band audio is the only available DTMF transport mechanism, the Cisco Unified Border Element can be used to translate the in-band audio DTMF signaling into RFC 2833 signaling.

Three DTMF options are available on Unified CM SIP trunks:

- DTMF Signaling Method: No Preference

  In this mode, Unified CM attempts to minimize the usage of MTP resources by selecting the most appropriate DTMF signaling method for the call.

  If both endpoints support RFC 2833 in-band DTMF, then no MTP is required.

  If both devices support an out-of-band DTMF mechanism, then Unified CM uses KPML over the SIP trunk.

  If both devices support both RFC 2833 In Band DTMF and Out of Band DTMF, then RFC 2833 is preferred.

  The only case where an MTP is required is when one of the endpoints supports only out-of-band DTMF and the other supports only RFC 2833 in-band DTMF.

  The majority of Cisco Collaboration System endpoints support both in-band and out-of-band DTMF.

- DTMF Signaling Method: RFC 2833

  By placing a restriction on the DTMF signaling method across the trunk, Unified CM is forced to allocate an MTP if any one or both of the endpoints do not support RFC 2833 in-band DTMF. In this configuration, the only time an MTP will not be allocated is when both endpoints support RFC 2833 in-band DTMF.

- DTMF Signaling Method: OOB and RFC 2833

  In this mode, the SIP trunk signals to use both out-of-band (OOB) DTMF (KPML or Unsolicited Notify) and RFC 2833 in-band DTMF across the trunk, and it is the most intensive MTP usage mode. The only cases where MTP resources will not be required is when both endpoints support both RFC 2833 in-band DTMF and out-of-band DTMF.

Cisco recommends configuring the DTMF Signaling Method to **No Preference** on Unified CM SIP trunks. This setting allows Unified CM to make an optimal decision for DTMF and to minimize MTP allocation.

Cisco Unified Border Element supports any or all of the following SIP-based DTMF Relay transport methods on VoIP dial peers: RFC 2833 (rtp-nte), Unsolicited Notify (sip-notify), and KPML (sip-kpml).

Codec Selection over SIP Trunks

Before media can be established between communicating entities, both the entities must agree on the codec(s) that they want to use. This codec (or codecs, if both audio and video are involved) is derived from the intersection of codecs supported by communicating entities involved and the configured policy in Unified CM, configured by *region* settings.

The region settings in Unified CM provide for configurable audio codec preference lists. In addition to the default Lossy and Low Loss audio codec preference lists that can be selected via a region's Link Loss Type, multiple custom audio codec preference lists can also be created. Audio codec preference lists can be used for codec selection for calls within a region and between regions. The Maximum Audio Bit Rate is still applied for calls within a region and between regions; but rather that using the highest audio quality codec (as in earlier Unified CM releases) based on the maximum bit rate setting, the codec selection is made based on the codec order in the audio codec preference list and the codecs that the endpoints support. (See Figure 6-17 and Figure 6-18.)

An Audio Codec Preference List is a list of all the codec types supported by Unified CM. The preference order of this list of codecs can be modified and saved as a custom preference list. (Note that codecs cannot be removed from the Audio Codec Preference List). The list of codecs used for codec negotiation during call setup is the subset of codecs supported by the device and those in the codec preference list, limited by the maximum audio bit rate for the region or region pair.

Figure 6-17 and Figure 6-18 show examples of how codecs are selected for codec negotiation during call setup.

Figure 6-17 Codec Selection with Maximum Audio Codec Bit Rate of 64 kbps

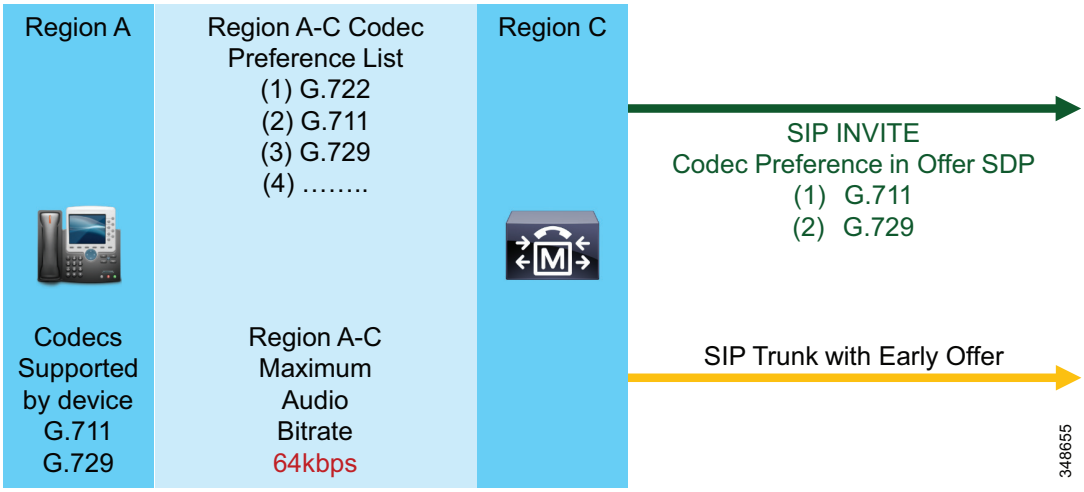
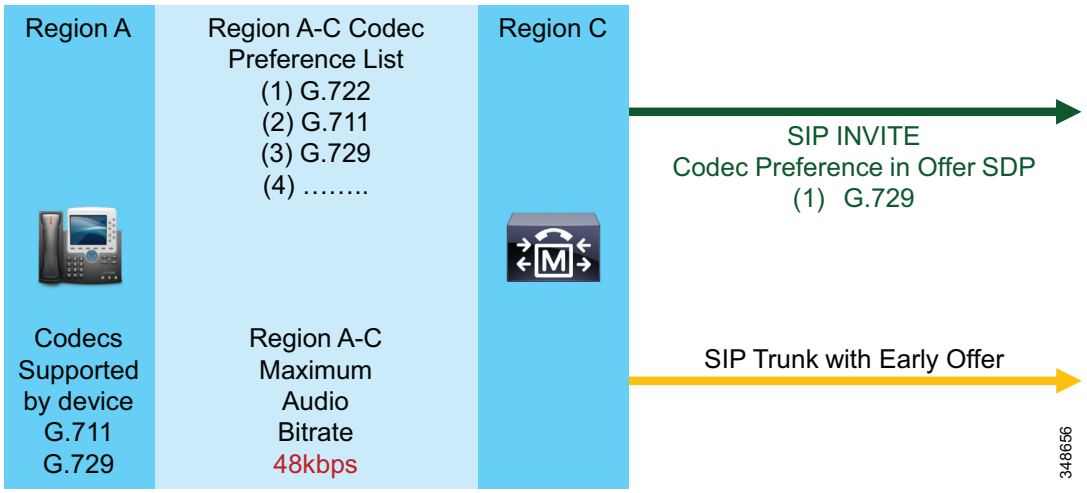
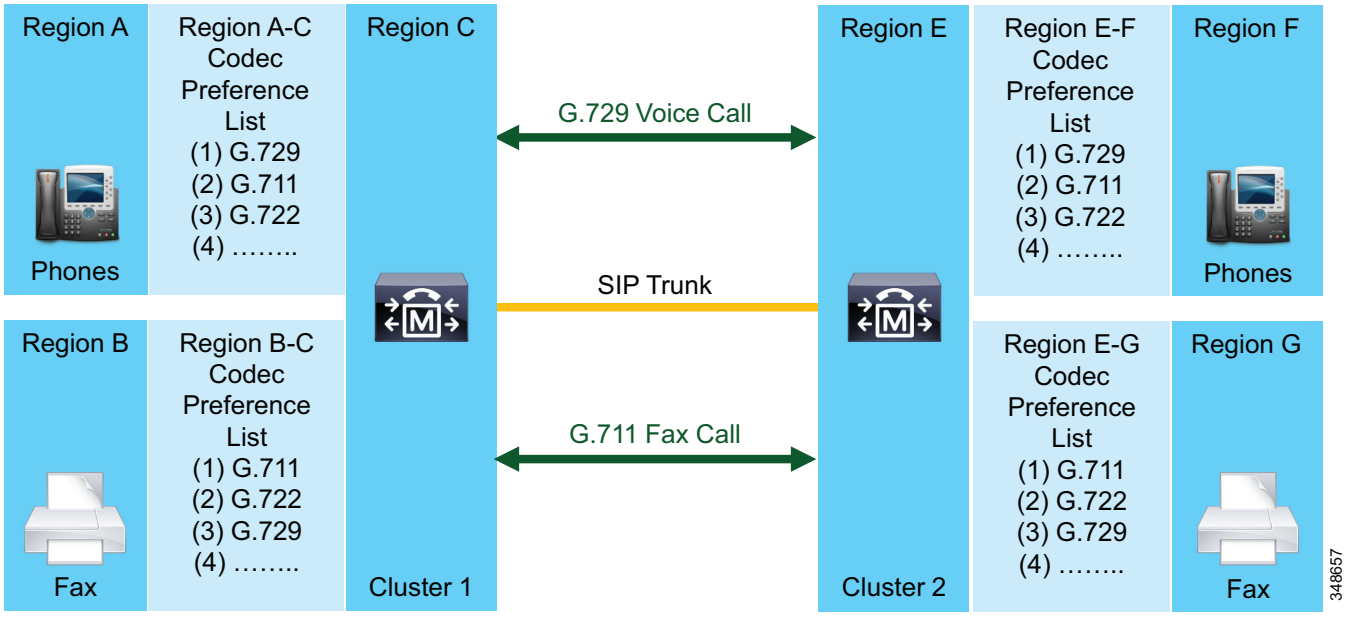


Figure 6-18 Codec Selection with Maximum Audio Codec Bit Rate of 48 kbps



For calls between two Unified CM clusters over SIP intercluster trunks, audio codec preference lists allow the codec to be selected for a call based upon the codec preferences of the calling and called devices. By grouping devices in each cluster into regions based on their codec preferences, a single intercluster trunk can be used to support multiple calls, with each call type using its preferred codec. (See Figure 6-19.)

Figure 6-19 Audio Codec Preference Lists for Voice and Fax Calls Between Two Unified CM Clusters



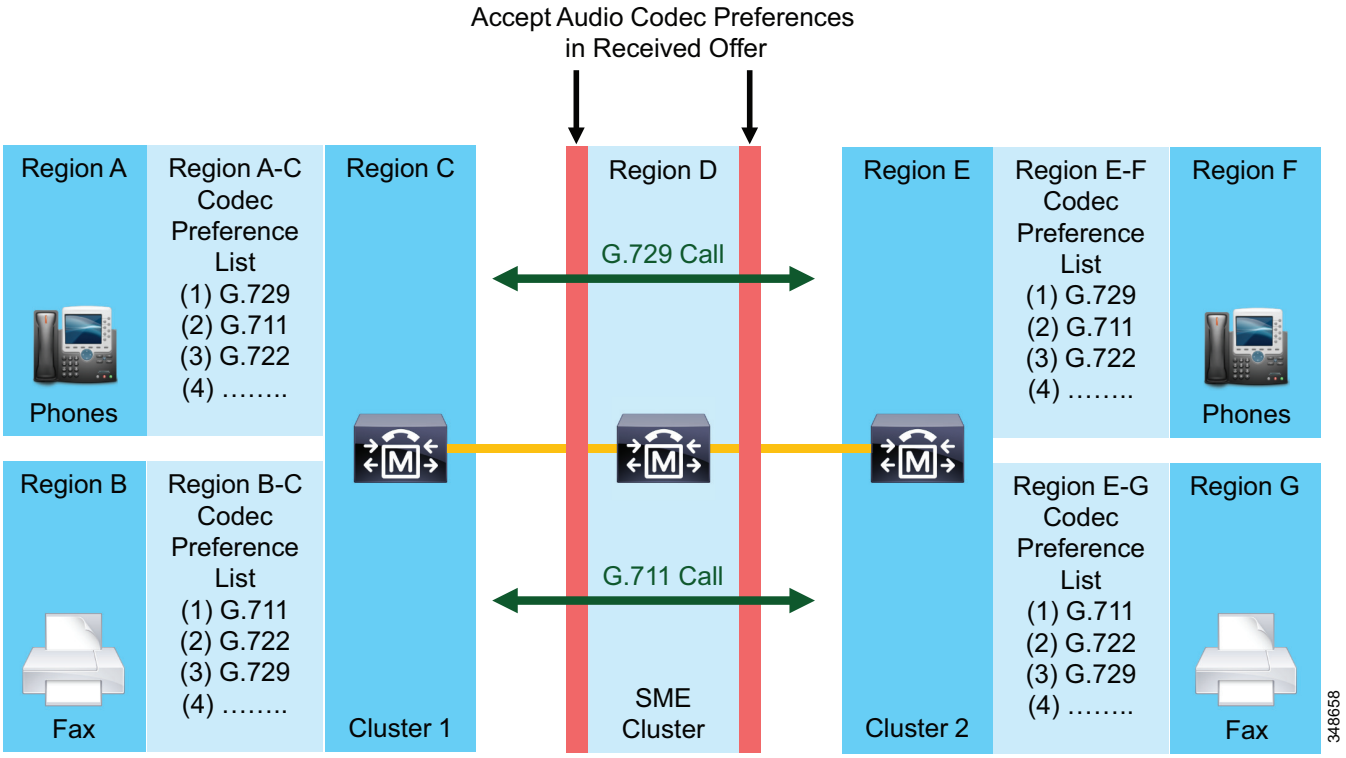
Note Configure equivalent inter-region audio codec preference lists for each device type in each cluster to ensure that a common codec is selected for each device type, irrespective of call direction or trunk configuration. If the audio codec preference lists in each cluster are not equivalent, the codecs used per call can vary based on call direction and trunk configuration. (Ordinarily, the codec preference order is not honoured by the cluster receiving the codec preference list.)

Note Do not use SIP trunks configured for Early Offer with **MTP Required** enabled if codec preference is required. This trunk configuration inserts an MTP for inbound and outbound calls, which is limited to a single audio codec only, thereby overriding codec preference and selection.

Accept Audio Codec Preferences in Received Offer

In deployments where calls can pass through more than one Unified CM cluster (for example, SME deployments), the inter-region audio codec preference list of the intermediary Unified CM (SME) cluster can override the preferred codec selection between the calling and called devices. To ensure that the endpoints' codec preferences are honoured as calls pass through SME, enable the SIP Profile feature **Accept Audio Codec Preferences in Received Offer** on all SME SIP trunks. (See Figure 6-20.)

Figure 6-20 SME Deployment Using "Accept Audio Codec Preferences in Received Offer" on SIP Trunks



Note The **Accept Audio Codec Preferences in Received Offer** feature is available only on SIP trunks (a SIP Profile feature). This feature does not offer consistent results if used in an SME deployment where the SME cluster uses a combination of SIP, H.323, and/or MGCP trunks. Therefore, the **Accept Audio Codec Preferences in Received Offe**r feature should be used when the SME cluster is deployed using only SIP trunks.

Cisco Unified CM and Cisco Unified Border Element SIP Trunk Codec Preference

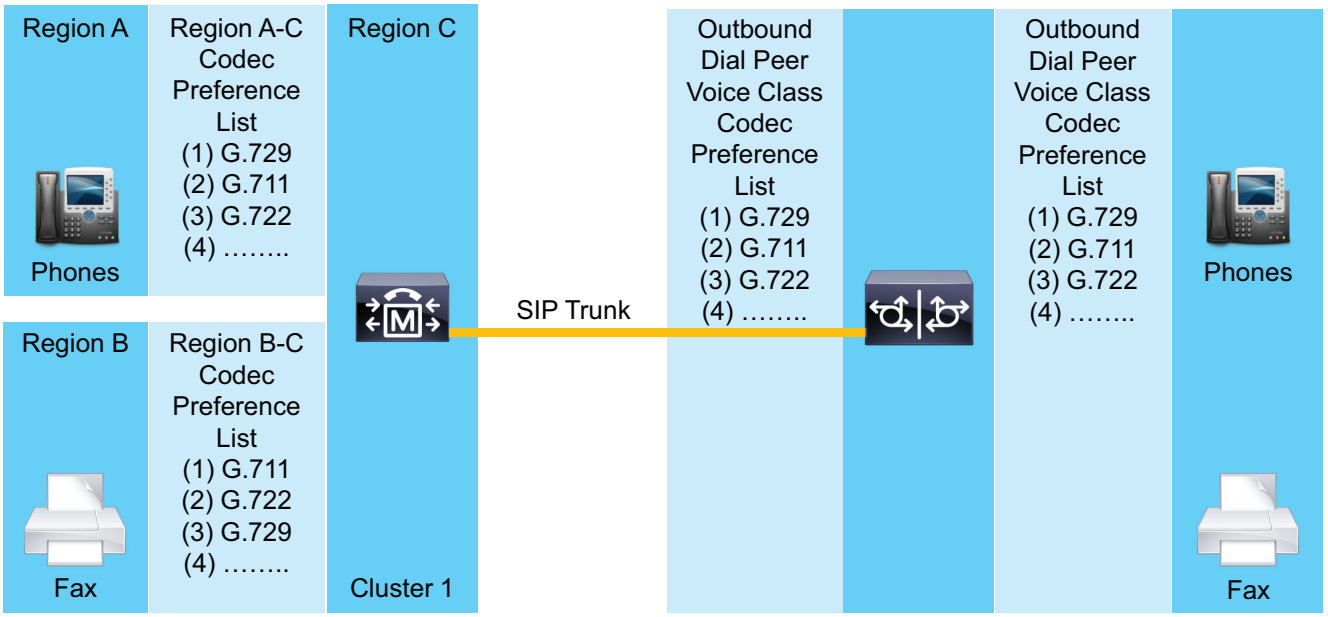
Unified CM audio codec preference lists can be used in Unified Communications deployments with Cisco Unified Border Element to simplify configuration of SIP trunks between Unified CM and Unified Border Element. For example, instead of using dedicated SIP trunks to the Unified Border Element for voice and fax calls, a single Unified CM SIP trunk can be used where the codec preference for each device type is honored as calls pass through the Unified Border Element.

In Figure 6-21 the Voice Class Codec Preference lists defined on Cisco Unified Border Element's inbound and outbound dial peers do not change the preference of the listed codecs in the received Offer. Cisco Unified Border Element does codec filtering on the received Offer, both on the inbound and outbound dial-peer, and passes across the common codecs in the same preference order as received in the inbound Offer to the peer leg.

If codecs, in addition to those received in an Offer, are defined in the voice class codec list, then these codecs will be appended to those received in the ordered list and sent out in the outbound Offer.

Thus, a single inbound and outbound dial-peer can be configured on Cisco Unified Border Element for all device types. Cisco recommends using the same voice class codec preference list for both the inbound and outbound dial-peer, with that list containing the codecs that you want to negotiate with the service provider. As mentioned above, the order of the codecs will be dictated first by the order received in the inbound Offer and then by the order defined in the voice class codec preference list.

Figure 6-21 Cisco Unified CM and Cisco Unified Border Element SIP Trunk Codec Preference



SIP Trunk Transport Protocols

SIP trunks can use TCP, TLS (which runs over TCP), or UDP as a message transport protocol. Unified CM provides a native interworking function for SIP trunks using different transport protocols. TCP is recommended within Cisco Collaboration Systems networks because it is a reliable and connection-orientated protocol with the capability to fragment and re-assemble large SIP messages. UDP is not connection-orientated or reliable (message delivery is not guaranteed), and it relies on the SIP Invite Retry count and SIP Trying timers to detect and respond to far-end device failures. Cisco recommends using SIP OPTIONS Ping to dynamically track the state of each destination IP address on each SIP trunk and the collective state of the trunk as a whole.

For more information on SIP trunk timer tuning, refer to the configuration example and technical notes at

https://www.cisco.com/en/US/products/sw/voicesw/ps556/products_configuration_example09186a008082d76a.shtml

Note Although TCP is the recommended transport protocol within a Cisco Collaboration Systems network, most service providers prefer to use UDP because it has a lower processing overhead than TCP. Cisco Unified Border Element can be used to provide TCP-based SIP trunk connections to the Cisco Collaboration Systems network and UDP-based SIP trunk connections to service provider networks.

Secure SIP Trunks

Securing SIP trunks involves two processes:

- Configuring the trunk to encrypt media (see Media Encryption, page 6-32)
- Configuring the trunk to encrypt signaling (see Signaling Encryption, page 6-32)

Media Encryption

Media encryption can be configured on SIP trunks by checking the trunk's **SRTP allowed** check box. It is important to understand that enabling **SRTP allowed** causes the media for calls to be encrypted, but the trunk signaling will not be encrypted and therefore the session keys used to establish the secure media stream will be sent unencrypted. It is therefore important that you ensure that signaling between Unified CM and its destination SIP trunk device is also encrypted so that keys and other security-related information do not get exposed during call negotiations.

Signaling Encryption

SIP trunks use TLS for signaling encryption. TLS is configured on the SIP Security Profile associated with the SIP trunk, and it uses X.509 certificate exchanges to authenticate trunk devices and to enable signaling encryption.

Certificates can be either of the following:

- Imported to each Unified CM node from every device that wishes to establish a TLS connection to that node's SIP trunk daemon
- Signed by a Certificate Authority (CA), in which case there is no need to import the certificates of the remote devices; only the CA certificate needs to be imported

Unified CM provides a bulk certificate import and export facility. However, for SIP trunks using **Run on all Active Unified CM Nodes** and up to 16 destination addresses, using a Certificate Authority provides a centralized and less administratively burdensome approach to setting up signaling encryption on SIP trunks.

For more information on TLS for SIP trunks, refer to the latest version of the *Cisco Unified Communications Manager Security Guide*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

For information on certificate authorities, refer to the Certificate Authority (CA) information in the latest version of the *Cisco Unified Communications Operating System Administration Guide*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

If the system can establish a secure media or signaling path and if the end devices support SRTP, the system uses an SRTP connection. If the system cannot establish a secure media or signaling path, or if at least one device does not support SRTP, the system uses an RTP connection. SRTP-to-RTP fall-back (and vice versa) may occur for transfers from a secure device to a non-secure device or for conferencing, transcoding, music on hold, and so on.

For SRTP-configured devices, Unified CM classifies a call as encrypted if the **SRTP Allowed** check box is checked for the device and if the SRTP capabilities for the devices are successfully negotiated for the call. If these criteria are not met, Unified CM classifies the call as non-secure. If the device is connected to a phone that can display security icons, the phone displays the lock icon when the call is encrypted.

Note MTPs that are statically assigned to a SIP trunk by means of the **MTP Required** checkbox do not support SRTP because they do not support the pass-through codec.

To ensure that SRTP is supported for all calls, configure the SIP trunk for Delayed Offer or **Best Effort Early Offer**.

Where **Early Offer support for voice and video calls Mandatory (insert MTP if needed)** is configured for devices that support encryption, all calls that do not need to use MTPs can support SRTP. When an MTP is inserted into the call path, this dynamically inserted MTP supports the pass-through codec, and encrypted calls are supported in the following cases:

- If the calling device is an older SCCP-based phone registered to Unified CM, SRTP can be negotiated in the initial call setup.
- If the call arrives inbound to Unified CM on a Delayed Offer SIP trunk or an H.323 Slow Start trunk, SRTP will not be negotiated in the initial call setup because no security keys are available, but the call can be escalated mid-call to support SRTP if the call media is renegotiated (for example, after hold or resume).

If Unified CM dynamically inserts an MTP for reasons other than Early Offer, such as for a Trusted Relay Point or as an RSVP agent, then SRTP will be supported with an MTP that supports the pass-through codec (Cisco IOS MTPs).

Note In-band to out-of-band DTMF conversion using MTPs does not function for SRTP encrypted media streams because the MTP is unable to decrypt the DTMF packets.

User Identity and SIP Trunks

A calling user's name and number can be sent over Unified CM SIP trunks in the following SIP message headers:

From:	From: "Jim Bob" <sip:1000@10.10.199.250> From: "Anonymous" <sip:localhost>
To:	To: "Nick Cave" <sip:2000@10.10.100.251>
P-Asserted-Identity:	P-Asserted-Identity: "Jim Bob" <sip:1000@10.10.199.250>
Remote-Party-ID:	Remote-Party-ID: "Jim Bob" <sip:1000@10.10.199.250>

The From and To message headers sent in SIP Requests and Responses indicate the direction of the call. (The From header represents the calling user and the To header represents the called user.) The From and To headers remain the same in all SIP Requests and Responses for the call.

SIP allows the From header to be made anonymous so that the calling user information is not presented to the called user.

The P-Asserted-Identity and Remote-Party-ID headers (if present) always contain the user's identity. The user information contained in SIP messages with these identity headers is directional, so that the headers contain the calling user's identity in an Initial INVITE and the called user's identity in Responses. The P-Asserted-Identity and Remote-Party-ID headers can be used to trace the identity of an anonymous call.

By default, both the P-Asserted-Identity and Remote-Party-ID headers are sent over Unified CM SIP trunks, but they can be disabled. The usage of P-Asserted-Identity and Remote-Party-ID headers will depend upon the device that the Unified CM SIP trunk is connected to. P-Asserted-Identity is a more recent standard and more commonly used than Remote-Party-ID. The P-Asserted-Identity standard (RFC 3325) is considered to be more secure than Remote-Party-ID because it supports authentication between untrusted SIP Realms. For SIP trunk connections to untrusted networks, configure Unified CM to send a P-Preferred-Identity header instead of a P-Asserted-Identity header. Unified CM will respond to a Digest authentication challenge for the sent the P-Preferred-Identity header.

Caller ID Presentation and Restriction

As discussed above the calling user's name and number can be anonymized in the From header in SIP messages sent over a SIP trunk. Calling name and number presentation and restriction can be enabled in three ways:

- By configuring calling name and calling number presentation or restriction in a translation pattern associated with the calling device
- By configuring calling name and calling number presentation or restriction on the Unified CM trunk
- By configuring the P-Asserted-Identity related, SIP Privacy value on the Unified CM SIP trunk

These caller ID presentation and restriction configuration options operate in the following precedence order (highest precedence first):

1. SIP Privacy value
2. Trunk configuration
3. Device configuration

Called and Calling Party Number Normalization and SIP Trunks

As calls traverse the edge between the public PTSN or IP PTSN and the private enterprise network, the called and calling party numbers sent in call setup messages should ideally be normalized to a globally routable international format such as +E.164. How and where these numbers are normalized depends upon the type of PSTN network to which your enterprise is connected:

ISDN and Q.931 PSTN Networks

Calls within ISDN and Q.931 PSTN networks provide additional information in the Number Type fields of call setup messages to classify called and calling numbers. Number-types can be one of four types: Unknown, Subscriber, National, or International. For calls from the PSTN to the enterprise network, the number-type parameter can be used by the enterprise to globalize the calling number to its +E.164 value by prefixing it with the appropriate digits. Using a globalized PSTN calling number within the enterprise allows calls to be returned to the PSTN caller with little or no additional digit manipulation. Depending on the number format sent by the service provider, the enterprise called number might also have to be modified to match that of the enterprise dial plan. Cisco recommends deploying a +E.164 dial plan within the enterprise.

For more details and examples on how these number-types are used and dial plan recommendations, refer to the chapter on Dial Plan, page 14-1.

SIP-Based IP PSTN Networks

Calls from SIP-based IP PSTN networks do not include number type information in SIP messages. In this case, the IP PSTN service provider should present the PSTN calling number using a globally routable international representation (for example, a +E.164 number). Depending on the number format sent by the service provider, the enterprise called number might have to be modified to match that of the enterprise dial plan. Cisco recommends deploying a +E.164 dial plan within the enterprise.

If the service provider sends the PSTN calling number in +E.164 format and the called number in a format that matches that used by the enterprise dial plan (+E.164 recommended), then little or no changes need to be made to these numbers within the enterprise.

The inability of SIP to transport the number type implies that the normalization of the calling number must be performed before the call is presented to Unified CM's call routing process. One place where the transformation can be performed is on the ingress SIP gateway. The following example configuration shows the translation rules that can be defined on a Cisco IOS gateway to accomplish this transformation:

```
voice translation-rule 1
 rule 1 // /+4940/ type subscriber subscriber
 rule 2 // /+49/ type national national
 rule 3 // /+/ type international international
...
voice translation-profile 1
 translate calling 1
...
dial-peer voice 300 voip
 translation-profile outgoing 1
 destination-pattern .T
 session protocol sipv2
 session target ipv4:9.6.3.12
...
```

When configured as in the example above, a Cisco IOS gateway using SIP to communicate with Unified CM will send calling party information digits normalized to the E.164 format, including the + sign. The Unified CM configuration will receive all calls from this gateway with a numbering type of "unknown" and will not need to add any prefixes.

For more details on configuring translation rules, refer to the *Voice Translation Rules* document, available at

https://www.cisco.com/en/US/tech/tk652/tk90/technologies_tech_note09186a0080325e8e.shtml

Unified CM can set the calling party number of outgoing calls to the normalized global format. The number-type in outgoing calls from the SIP trunk will be "unknown," and the Cisco IOS gateway should change it to International if no stripping is done, or perform a combination of stripping and numbering type change if required by the connected service provider.

Reasons for Using Only SIP Trunks in Cisco Collaboration Systems Deployments

For Cisco Collaboration Systems networks consisting of one or more Unified CM clusters, Unified Communications applications, Session Border Controllers, and gateways, using SIP as the sole interconnecting trunk protocol allows you to build a simplified Collaboration Systems network with a rich set of common features.

When compared to other trunk protocols, SIP trunks today support a number of unique features, such as:

- SIP OPTIONS Ping which tracks the overall operational status of the trunk and the state of each trunk's destination nodes.
- Codec Preference Lists and the ability to accept the codec preferences received in an SDP Offer.
- Support for H.264 video with BFCP-based presentation sharing and Far End Camera Control.
- SIP message normalization and transparency, which provides powerful script-based functionality for SIP trunks that can be used to transparently forward and/or modify SIP messages and message body (SDP) contents as they traverse Unified CM. Normalization and transparency scripts are designed to address SIP interoperability issues, allowing Unified CM to interoperate with SIP-based third-party PBXs, applications, and IP PSTN services.
- Support for IPv4 only, IPv6 only, or Dual Stack (IPv4 and IPv6) ANAT-enabled SIP trunks.

Design and Configuration Recommendations for SIP Trunks

When designing and deploying a SIP-based Cisco Collaboration Systems network, Cisco recommends that you use the following SIP trunk features:

Best Effort Early Offer for Unified CM Leaf Clusters and SME Clusters

Using only SIP trunks configured as **Best Effort Early Offer**, eliminates MTP usage for Early Offer creation in leaf clusters and makes SME clusters transparent to media negotiation. With **Best Effort Early Offer**, the SIP trunk sends an Early Offer only if it has enough information about the calling device's media capabilities to create the Offer; if it does not have this information, it sends a Delayed Offer instead.

Prior to **Best Effort Early Offer**, the decision to use Delay Offer or Early Offer on leaf cluster trunks was typically based upon the number of older SCCP endpoints registered to the cluster. Because older SCCP endpoints require the insertion of an MTP to create an Offer for calls over Early Offer SIP trunks, where large numbers of SCCP endpoints exist within the cluster, Delayed Offer was preferred to avoid MTP usage. **Best Effort Early Offer** removes the need to decide upon Early Offer or Delayed Offer SIP trunk configuration based on the type of endpoints registered to the cluster.

In Cisco Collaboration System deployments, receipt of Early Offer only may be required by non-Cisco Unified Communications applications and services. There are two options to address the requirement that Early Offer is always received:

- Cisco Unified Border Element provides a SIP Delayed Offer to Early Offer feature for voice calls, which converts inbound Delayed Offer calls to outbound Early Offer calls, thus allowing Unified CM and SME to use **Best Effort Early Offer** trunks. A typical example of this use case is service provider IP PSTN connections, which typically must always receive SIP Early Offer.
- For enterprise Unified Communications applications that accept only SIP Early Offer, a dedicated Early Offer SIP trunk can be used from the Unified CM Leaf cluster to the Unified Communications application. If a large number of MTPs are required on the Early Offer SIP trunk, consider using the Cisco Unified Border Element Delayed Offer to Early Offer conversion feature.

In SME clusters, **Best Effort Early Offer** performs the same role as **MTP-Less Early Offer** by making the SME cluster transparent to media negotiation, which in turn forces media decisions to be made by the end Unified Communications system where, if required, media resources can be inserted to address DTMF or codec mismatch issues. Media resources must not be associated with MTP-Less Early Offer SME trunks. If needed, media resources can be associated with **Best Effort Early Offer** trunks.

Run on All Unified CM Nodes

This feature is supported on SIP trunks and route lists, and it greatly simplifies call routing from and through Unified CM and SME clusters. Cisco highly recommends enabling the **Run on all Unified CM nodes** feature on all SIP trunks and route lists. Call routing is simplified through a combination of the **Run on all Unified CM nodes** and the Route Local features, whereby phone calls over SIP trunks will always originate from the Unified CM node where the phone is registered. Likewise for trunk to trunk calls, the outbound SIP trunk call will always originate from the Unified CM node on which the inbound Trunk call arrived. Enabling **Run on all Unified CM nodes** on all SIP trunks and route lists eliminates the need to set up calls between call processing nodes within the cluster, which can be useful when clustering over the WAN is deployed within a Unified CM or SME cluster.

Up to 16 SIP Trunk Destination IP Addresses

SIP trunks can be configured with up to 16 destination IP addresses, 16 fully qualified domain names, or a single DNS SRV entry. Support for additional destination IP addresses reduces the need to create multiple trunks associated with route lists and route groups for call distribution between two Unified Communications systems, thus simplifying Unified CM trunk design. When IP addresses are used as destinations on a SIP trunk, Unified CM randomly distributes calls across all defined destination IP addresses.

SIP OPTIONS Ping

Enable the SIP OPTIONS Ping feature on the SIP Profile associated with a SIP trunk to dynamically track the state of each the trunk's destinations and the overall state of the trunk.

PRACK

PRACK provides reliability of 1XX responses for interoperability scenarios with the PSTN, and it can also be used to reduce the number of SIP messages that need to be exchanged before setting up two-way media. Enable PRACK through the **SIP Rel1XX Options** parameter on the SIP Profile associated with the trunk.

SIP Trunk DTMF Signaling Method – No Preference

Using **DTMF Signaling Method: No Preference** is recommended on SIP trunks. In this mode Unified CM attempts to minimize the usage of MTP resources by selecting the most appropriate DTMF signaling method (in-band or out-of-band) for the call.

Unified CM Session Management Edition

Cisco Unified Communications Manager Session Management Edition (Unified CM SME) is the recommended trunk and dial plan aggregation platform in multi-site distributed call processing deployments. SME is essentially a Unified CM cluster with trunk interfaces only and no IP endpoints. It enables aggregation of multiple unified communications systems, referred to as leaf systems. (See Figure 6-22.)

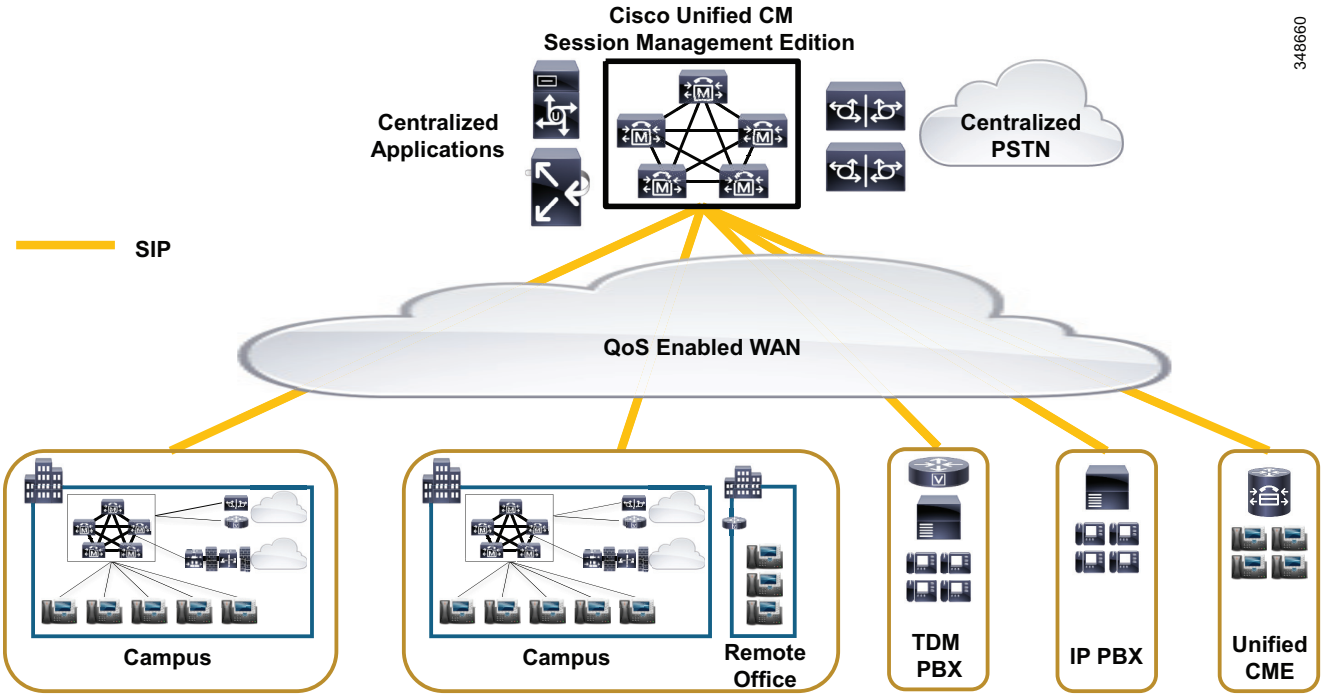
SIP trunks are highly recommended for SME and Leaf Unified Communications systems because SIP offers additional features and functionality not available in H.323 and MGCP trunks. As discussed later in this section, there are certain trunk features that are exclusive to SME designs that use SIP trunks only. If your Unified Communications network must support H.323 or MGCP trunk connections to gateways or other Unified Communications applications, to preserve the SIP-only trunk features in your SME cluster, connect these H.323 and/or MGCP trunks to leaf Unified Communications systems instead of SME.

Cisco Unified CM Session Management Edition (SME) supports the following call types:

- Voice calls
- Video calls
- Encrypted calls
- Fax calls

Unified CM Session Management Edition may also be used to connect to the PSTN and third-party unified communications systems such as PBXs and centralized unified communications applications.

Figure 6-22 Multisite Distributed Call Processing Deployment with Unified CM Session Management Edition



When to Deploy Unified CM Session Management Edition

Cisco recommends deploying Unified CM Session Management Edition if you want to do any of the following:

- Create and manage a centralized dial plan

  Rather than configuring each unified communications system with a separate dial plan and trunks to connect to all the other unified communications systems, Unified CM Session Management Edition allows you to configure the leaf unified communications systems with a simplified dial plan and trunk(s) pointing to the SME cluster. Unified CM Session Management Edition holds the centralized dial plan and corresponding reachability information about all the other unified communications systems.

  Note Running ILS GDPR on SME and Unified CM leaf clusters further simplifies dial plan administration because individual directory numbers, E.164 numbers corresponding to DNs, route patterns (for internal and external number ranges), and URIs can be distributed using the ILS service. This approach simplifies dial plan administration by reducing the required number of route patterns to one SIP route pattern per call control system (Unified CM cluster, for example), instead of a route pattern for each unique number range. For more information on ILS and GDPR, see Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR), page 10-32.

- Provide centralized PSTN access

  Unified CM Session Management Edition can be used to aggregate PSTN access to one (or more) centralized PSTN trunks. Centralized PSTN access is commonly combined with the reduction or elimination of branch-based PSTN circuits.

- Centralize applications

  The deployment of Unified CM Session Management Edition enables commonly used applications such as conferencing or voice mail to connect directly to the SME cluster, thus reducing the overhead of managing multiple trunks to leaf systems.

- Aggregate PBXs for migration to a Unified Communications system

  Unified CM Session Management Edition can provide an aggregation point for multiple PBXs as part of the migration from legacy PBXs to a Cisco Unified Communications System. If ILS GDPR is deployed, the number ranges and/or URIs supported by each third party system can also be imported into ILS GDPR and reached by means of a SIP route pattern and corresponding SIP trunk.

Differences Between Unified CM Session Management Edition and Standard Unified CM Clusters

The Unified CM Session Management Edition software is exactly the same as Unified CM. However, Unified CM software has been enhanced to satisfy the requirements of this new deployment model. Unified CM Session Management Edition is designed to support a large number of trunk-to-trunk connections, and as such it is subject to the following design considerations:

Capacity and Sizing

It is important to size the Unified CM Session Management cluster correctly based on the expected BHCA traffic load between leaf Unified Communications systems (for example, between Unified CM clusters and PBXs), to and from any centralized PSTN connections, and to any centralized applications. Determine the average BHCA and call holding time for users of your Unified Communications system, and share this information with your Cisco account Systems Engineer (SE) or Cisco Partner to size your Unified CM Session Management Edition cluster correctly. For more information on SME sizing, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

SME Trunks

Although SME supports SIP, H.323, and MGCP trunks, Cisco highly recommends SIP as the trunk protocol of choice for SME and Unified CM leaf clusters running Cisco Unified Communications System Release 8.5 and later versions.

Using only SIP trunks in the SME cluster allows you deploy a "media transparent" cluster where media resources (when required) are inserted by the end or leaf Unified Communications system and never by SME. Using only SIP trunks also allows you use extended round-trip times (RTTs) between SME nodes when clustering over the WAN.

SME SIP trunks should be configured as **Best Effort Early Offer** trunks. Leaf Unified CM cluster SIP trunks should also be configured as **Best Effort Early Offer**.

SME Transparency for Media Negotiation

When a media resource such as an MTP or transcoder is needed to allow a call to proceed successfully, these resources should be allocated by the edge or leaf Unified Communications systems. If SME trunk media resources are used for a call traversing the SME cluster, the media path call will hairpin through the SME media resource. By using SIP trunks only and **Best Effort Early Offer** (or **MTP-less Early Offer**), an SME cluster can be deployed without media resources. If or when media resources are required, they can be provided by the edge or leaf Unified Communications system.

Clustering over the WAN with SME CoW+

With Cisco Unified CM 9.1 and later releases, SME deployments support round-trip times (RTTs) of up to 500 ms between SME cluster nodes. (See Figure 6-23.) This extended RTT applies only to SME clusters (80 ms is the maximum RTT for standard Unified CM cluster designs) and is subject to the following design restrictions:

- SME deployments with extended clustering over the WAN (CoW+) round-trip times are supported with SIP trunks only. All SIP trunks must be configured as either all **Best Effort Early Offer** (preferred) or all **MTP-less Early Offer** and must use the **Run on all Unified CM Nodes** feature so that calls are not routed between nodes within the SME cluster. H.323, MGCP, and SCCP protocols are not supported for SME deployments with extended clustering over the WAN round-trip times.
- No endpoints or CTI devices are configured or registered to the SME cluster.

- No media resources such as MTPs, trusted relay points (TRPs), RSVP agents, or transcoders are configured or registered to the SME cluster. (To disable media resources hosted on Unified CM nodes, deactivate the IPVMS service on each node within the cluster.)
- A minimum of 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) traffic between sites.
- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, a minimum of 1.544 Mbps (T1) bandwidth is required for database and other inter-server traffic between the publisher node and every remote subscriber node.

Note As with all SME designs, your SME design must be reviewed and approved by the Cisco SME Team prior to deployment.

The upgrade process for an SME cluster consists of two key parts:

- Version switch-over — The call processing node is rebooted and initialized with the new software version (this takes approximately 45 minutes per server).
- Database replication — The subscriber's database is synchronized with that of the publisher node.

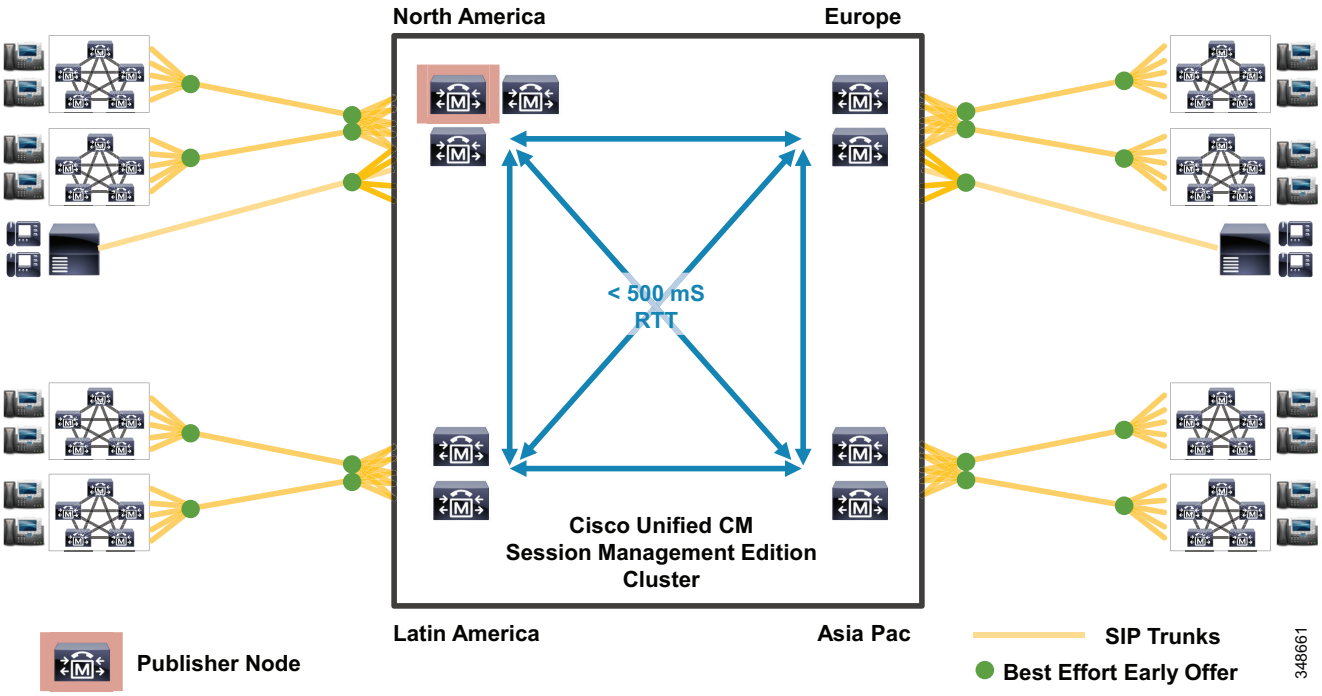
The time taken to complete this database replication phase depends on the number of subscribers nodes in the cluster and the RTT between the publisher and subscriber nodes. The database replication process has a minimal impact of the subscriber's call processing capability and typically can be run as a background process during normal SME cluster operation. Avoid making changes to the SME cluster configuration during the database replication phase because this increases the time it takes to complete the replication.

For SME clusters deployed with extended RTTs, before upgrading the cluster, run the following administrator-level CLI command on the publisher node:

utils dbreplication setprocess 40

This command improves replication setup performance and reduce database replication times.

Figure 6-23 Unified CM Session Management Edition – Clustering over the WAN with Extended Round Trip Times



Unified CM Versions

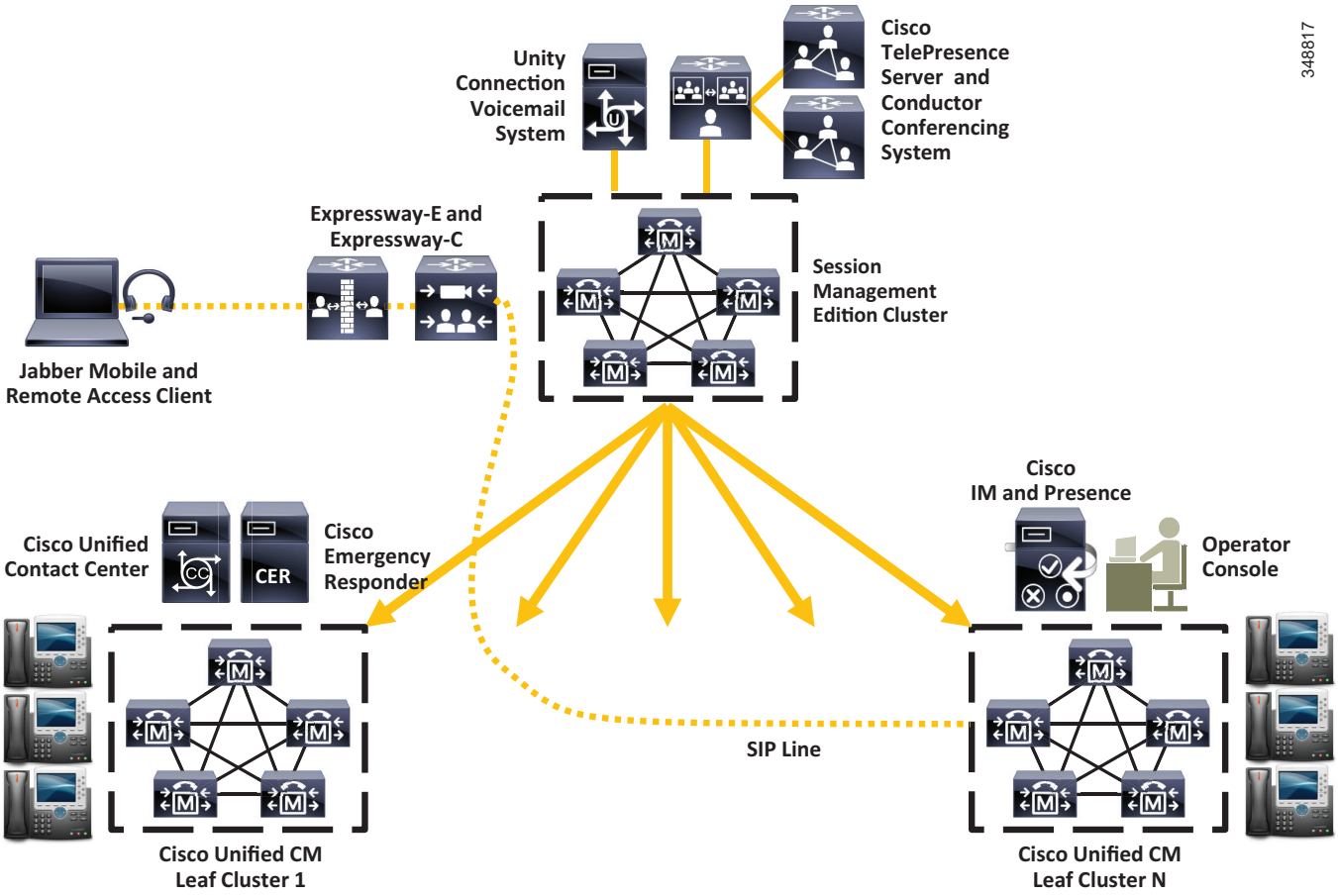
Using the latest Cisco Collaboration Systems Release and SIP trunks across all Unified CM leaf clusters and the SME cluster enables your deployment to benefit from common cross-cluster features such as codec preference lists, ILS, GDPR, and Enhanced Locations call admission control (CAC). If you do not wish to upgrade to the latest Unified CM version on all clusters, the lowest recommended version is Cisco Unified CM 8.5 using SIP trunks because this version includes features that improve and simplify call routing through Unified CM and Session Management Edition clusters.

Guidance on Centralizing Unified Communications Applications with Session Management Edition

Co-locating and connecting Unified Communications (UC) applications to the Session Management Edition cluster can provide economies of scale and a reduction in administrative overhead by using one centralized application instance rather than multiple instances associated with each leaf UC system. The following section provides some design guidance for UC applications the can be co-located with the Session Management Edition cluster.

As a general rule, applications that rely only on number-based call routing to establish calls can connect to Unified CM Session Management Edition. Applications that require additional interfaces (for example, CTI) to track device state (for example, Unified Contact Center and Attendant Console) must connect to the leaf cluster. (See Figure 6-24.)

Figure 6-24 Centralized UC Applications and Session Management Edition



Centralized Voice Mail – Unity Connection

Voicemail applications such as Cisco Unity Connection can be connected to the SME cluster and provide voicemail service to users on all leaf UC systems.

On the intercluster trunks between leaf clusters and SME, and on the trunk connections to the voicemail application, ensure that the original called party or redirecting number is sent with calls routed to voicemail.

For non-QSIG-enabled trunks, the original called party or redirecting number transport can be enabled by:

- Enabling inbound and outbound **Redirecting Number IE Delivery** on MGCP gateways, H.323 gateways, and H.323 trunks
- Enabling inbound and outbound **Redirecting Diversion Header Delivery** on SIP trunks

For QSIG-enabled SIP, MGCP, and H323 trunks, the original called party number is sent in QSIG Diverting Leg Information APDUs. The diversion information sent in QSIG APDUs over QSIG-enabled trunks does not pick up any calling party modifications and also does not honor the Voice Mail Box Mask setting. QSIG diversion information sent by Unified CM is always set to the redirecting DN without applying any transformations.

If the redirecting DN is configured as +E.164, the leading "+" is removed and the QSIG diversion information carries only the E.164 number without the "+" character.

Considerations for all QSIG Trunk Types

Using QSIG in UC networks today provides a limited number of feature benefits and generally is not recommended. The primary reason for using QSIG is to provide the Call Back feature. (As an alternative, Collaboration users can track another user's state by using presence information provided by the Cisco IM and Presence service.) If you do enable QSIG on trunks from leaf UC systems to SME, you should also enable QSIG on all intercluster trunks; this avoids a poor end user experience where a phone user finds that Call Back works for some (QSIG enabled) called users but not others.

On H.323, MGCP, and SIP trunks with QSIG tunneling enabled, all number information (including calling, called, and redirecting number information) is always taken from the encapsulated QSIG message and not from the outer H.323 message or SIP headers. This QSIG trunk operation can require specialized design considerations for a voicemail system centralized on SME and serving multiple leaf systems.

As a general recommendation, to enable a smooth end-to-end QSIG implementation, a uniform globally unique dial plan should be implemented across all UC systems.

If QSIG trunks are used, redirecting numbers cannot be normalized before they are sent to the centralized voicemail system. This limitation requires that the centralized voicemail system mailbox number for users in each leaf UC system should correspond to the number format of the directory numbers used in each leaf system. For example:

- Users with directory numbers in E.164 format should have a corresponding voicemail system mailbox number using the same E.164 format.
- Users with directory numbers in +E164 format should have a corresponding voicemail system mailbox number using the same E164 format and an alternate voice mailbox number using the +E.164 format.

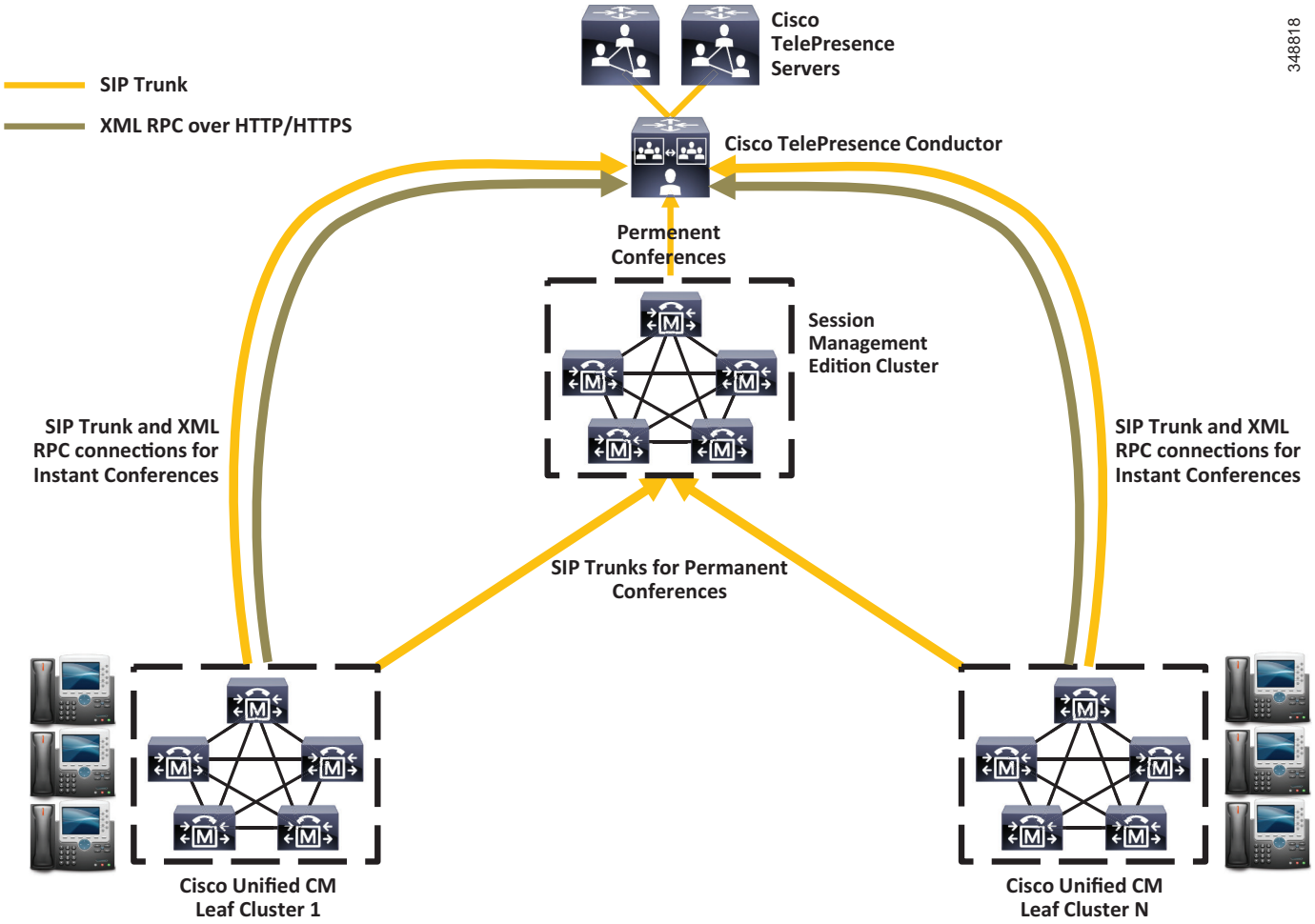
TelePresence Server and TelePresence Conductor

Conferencing systems can be connected to a Session Management Edition cluster. For deployments using Cisco TelePresence Conductor and TelePresence Servers, bear in mind that additional signaling connections beyond those of a SIP trunk are required for instant conferences. Unlike permanent conferences, which use route patterns and SIP trunks to reach their conferencing resource, Unified CM defines an instant conference as a media resource and uses XML RPC over HTTP/HTTPS to instruct TelePresence Conductor or the TelePresence Server to create the instant conference when a phone user presses the "conference" button. For instant conferences, HTTP/HTTPS XML RPC messages and SIP INVITE messages must come from the same source IP address, and therefore the instant conference connections (HTTP XML RPC and SIP trunk) must be configured on the leaf Unified CM cluster rather than in the SME cluster. TelePresence Conductor and TelePresence Servers can still be co-located with SME, but only permanent conferences can use a SIP trunk connection directly from the SME cluster, Instant conference SIP trunk and HTTP XML RPC connections must come directly from the leaf Unified CM cluster. (See Figure 6-25.)

For more information, refer to the latest version of the *Cisco TelePresence Conductor with Unified CM Deployment Guide*, available at

https://www.cisco.com/c/en/us/support/conferencing/telepresence-conductor/products-installation-and-configuration-guides-list.html

Figure 6-25 Centralized TelePresence Server and TelePresence Conductor



Expressway-C and Expressway-E

The Cisco Expressway platforms can be connected to a Session Management Edition cluster (see Figure 6-26). Depending on the deployment type, Expressway-C may or may not use a SIP trunk to connect to SME:

- Mobile and Remote Access

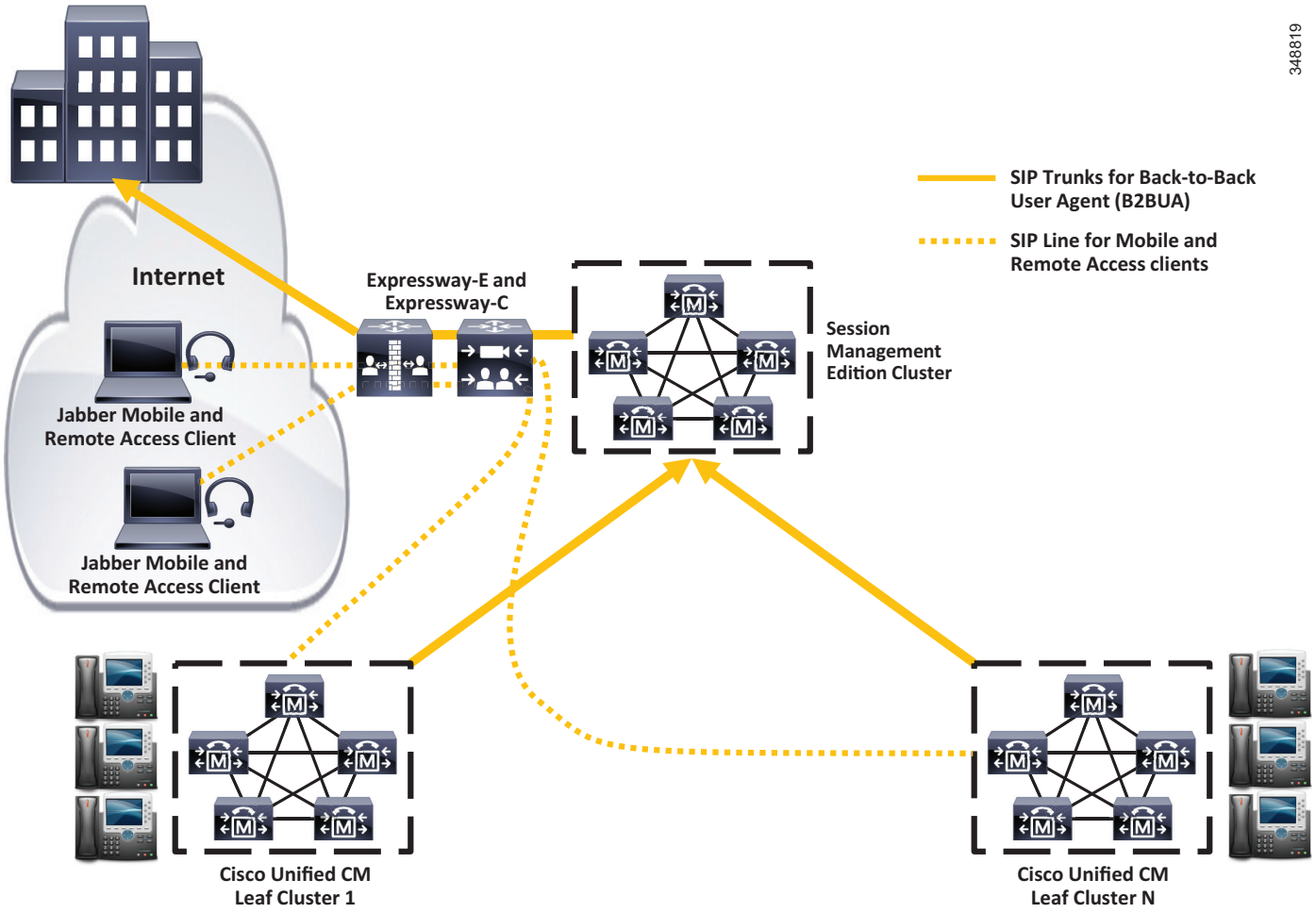
  Devices using the Mobile and Remote Access feature to connect to the enterprise UC network, do not establish a SIP connection to the SME cluster. The device uses UDS to discover and register directly to its home cluster. If SME is used to receive the UDS home cluster look-up requests, it also needs to use the ILS service to communicate with other Unified CM clusters for home cluster discovery. For more information, refer to the latest version of the *Mobile and Remote Access via Cisco Expressway Deployment Guide*, available at

  https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html

- Trunked Expressway Applications

  For Expressway application such Business-to-Business Collaboration, Expressway uses a direct SIP trunk connection to the SME cluster.

Figure 6-26 Centralized Expressway-C and Expressway-E



Summary of SIP Trunk Recommendations for Multi-Cluster SME Deployments

This section provides a summary of the SIP trunk recommendations and operation for multi- cluster deployments with Unified CM Session Management Edition.

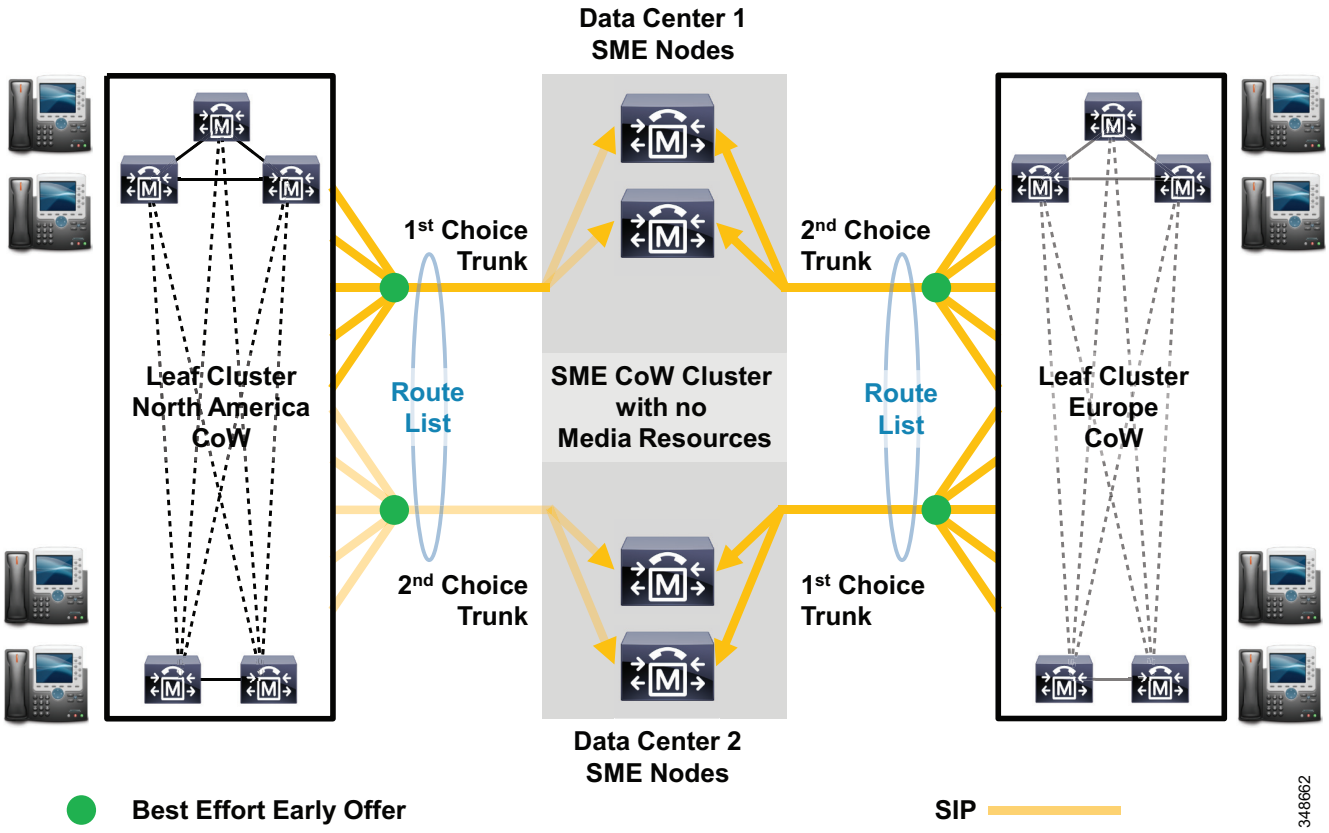
Recommendations for Unified CM Leaf Clusters:

- Configure one SIP trunk to each set of SME nodes in each regional data center. For example, if there are four regional SME data centers, create four SIP trunks in each leaf cluster (see Figure 6-27). This allows calls from all SME nodes to be received and accepted by the leaf clusters. Enable **Run on all Unified CM nodes** on all of these trunks.
- Place two or more of these leaf cluster SIP trunks into a route list and route groups for path redundancy to the SME CoW+ cluster.
- **Best Effort Early Offer** is recommended for all leaf cluster SIP trunks.

  In Unified Communications deployments, receipt of Early Offer only might be required by non-Cisco Unified Communications applications and services. For leaf clusters, there are two options to address the requirement that Early Offer is always received:

  - Cisco Unified Border Element provides a SIP Delayed Offer to Early Offer feature for voice calls, which converts inbound Delayed Offer calls to outbound Early Offer calls, thus allowing Unified CM and SME to use **Best Effort Early Offer** trunks. A typical example of this use case is for service provider IP PSTN connections via Cisco Unified Border Element, which typically must always receive SIP Early Offer.
  - For enterprise Unified Communications applications that accept only SIP Early Offer, a dedicated Early Offer SIP trunk can be used from the Unified CM Leaf cluster to the Unified Communications application. If a large number of MTPs are required on the Early Offer SIP trunk, consider using the Cisco Unified Border Element Delayed Offer to Early Offer conversion feature instead.
- Enable the IPVMS service on all leaf cluster nodes. Activate conferencing, music on hold, and annunciator resources as required. (Deactivating IPVMS-based MTPs is recommended.)
- As required, configure and associate Cisco IOS media resources (MTPs, conferencing, and transcoding) with the leaf cluster.
- Configure SIP trunk DTMF settings to **No Preference** (the default setting).
- Enable SIP Options Ping and PRACK.
- If required, configure and apply codec preference lists.

Figure 6-27 Recommended Trunk Configuration for CoW Leaf Cluster Trunks



Recommendations for Session Management Edition Clusters:

- Use only SIP trunks on the SME cluster.
- Configure one SIP trunk from the SME cluster to each leaf cluster (see Figure 6-28). Enable **Run on all Unified CM nodes** on these trunks, and configure trunk destinations to every call processing node in the leaf clusters.
- **Best Effort Early Offer** is recommended for all SME cluster SIP trunks.

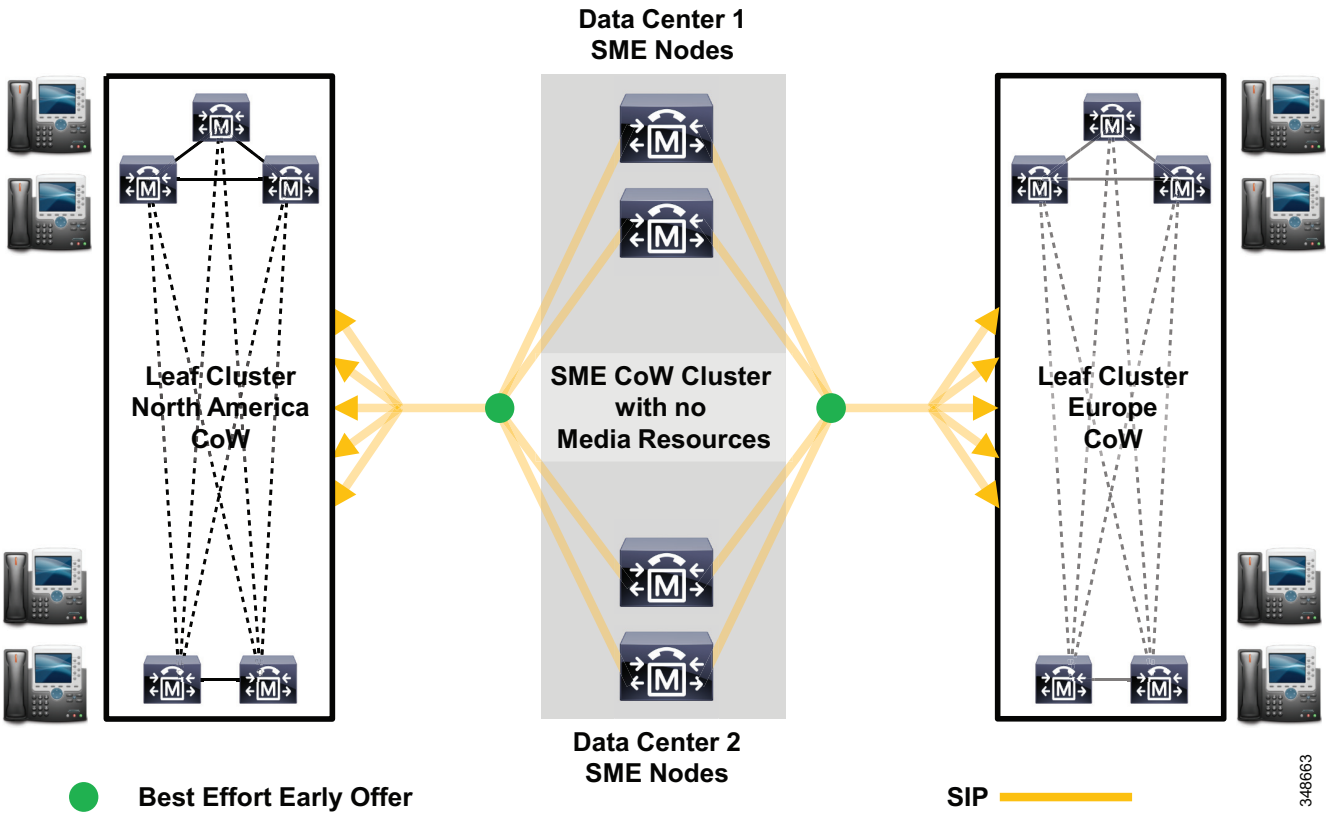
  If receipt of Early Offer only is required by non-Cisco Unified Communications applications and services connected to SME clusters, there are two options to address the requirement:

  - Cisco Unified Border Element provides a SIP Delayed Offer to Early Offer feature for voice calls, which converts inbound Delayed Offer calls to outbound Early Offer calls, thus allowing Unified CM and SME to use **Best Effort Early Offer** trunks only. A typical example of this use case is for service provider IP PSTN connections via Cisco Unified Border Element, which typically must always receive SIP Early Offer.
  - For enterprise Unified Communications applications that accept only SIP Early Offer, if a dedicated Early Offer SIP trunk is used from the SME cluster to the Unified Communications application, media resources will have to be associated with the SME trunks, which if used will cause unwanted media hair-pinning. The media resources typically used in this case are MTPs to create an Early Offer or address DTMF mismatches and transcoders to address codec mismatches. Using media resources in the SME cluster is not generally recommended; as an

alternative, consider using the Cisco Unified Border Element Delayed Offer to Early Offer feature between SME and the Unified Communications application, or use a direct trunk to the application from the leaf cluster.

- Disable the IPVMS service on all SME nodes. This disables Unified CM media termination points, conferencing, music on hold, and annunciator resources.
- Do not associate any Cisco IOS media resources with the SME cluster.
- Configure SIP trunk DTMF settings to **No Preference** (the default setting).
- Enable **Accept Audio Codec Preference in Received Offer** on all SME SIP trunks.
- Enable SIP Options Ping and PRACK.

Figure 6-28 Recommended Trunk Configuration for CoW+ SME Cluster Trunks



Recommendations for Call Routing Through Leaf and SME Clusters:

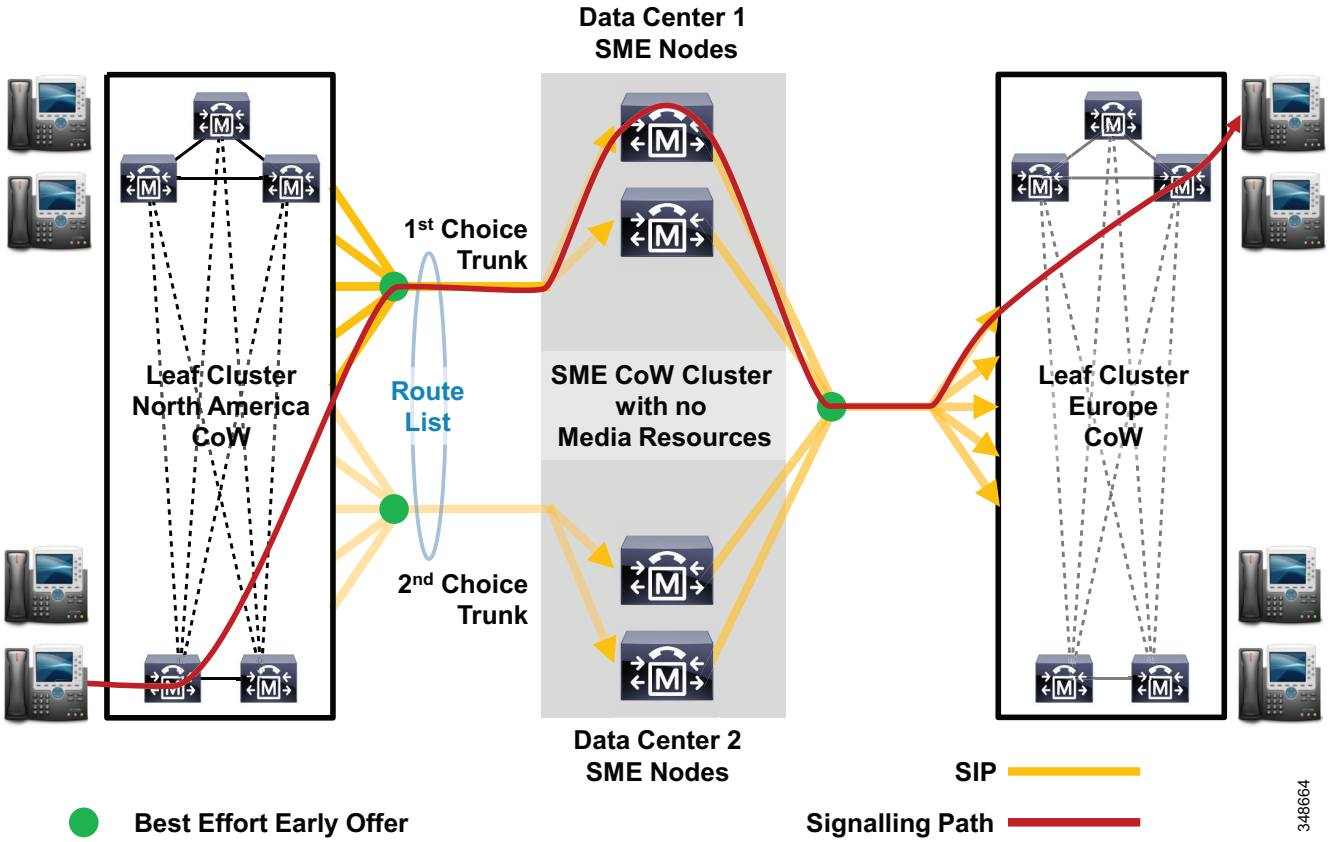
The outbound leaf cluster will originate a SIP trunk call from the same node that the calling device is registered to (using the Route Local rule). The leaf cluster will randomly select a destination address from the SIP trunk route list. (For the example in Figure 6-29, the first-choice trunk is selected.)

Outbound calls from the SME cluster will originate from the same node that the inbound call arrived on (using the Route Local rule). With **Run on all Unified CM nodes** enabled on all SME trunks, calls will never be set up between call processing nodes within the SME cluster. The SME cluster will randomly select a destination address on the SIP trunk pointing toward the destination leaf cluster.

For inbound SIP trunk calls to the destination leaf cluster, calls may be extended from the call processing node on which the inbound call arrived, to the node where the called device is registered.

Media resources, if needed, will be inserted by the leaf clusters (or end Unified Communications systems). If the device in the calling leaf cluster SIP trunk uses Delayed Offer, the media decision will be made by this cluster, which will insert media resources (MTPs and/or transcoders) as required. If the device in the calling leaf cluster SIP trunk send an Early Offer, the media decision will be made by the destination leaf cluster, which will insert media resources (MTPs and/or transcoders) as required.

Figure 6-29 Recommended Trunk Configuration for Call Routing Through Leaf and SME Clusters



Minor Features of Unified CM SIP Trunks

This section describes the function and usage of several minor features available on Unified CM SIP trunks.

Send sendrecv in Mid-Call INVITE

This feature is used to address interoperability issues with third-party products. When Unified CM places a call hold over a SIP trunk, it sends a mid-call INVITE with audio direction media attribute a=inactive in the SDP body to disconnect the media connection. On call resumption, Unified CM sends a Delayed Offer INVITE (without SDP) to the held device to obtain its media characteristics through an SDP Offer. According to RFC 3261 (section 14.2), the held device should construct the Offer as if it

were making a new call; that is, with a list of all supported codecs and a=sendrecv. Some third-party products respond with only the last used codec and media direction attributes, with the result that the call always remains in the inactive state and media cannot be resumed. When **Send "sendrecv" in mid call INVITE** is enabled, this feature inserts an MTP into the media path between the calling and calling devices, allowing the media connection to be disconnected between the Unified CM device and the MTP, while establishing and maintaining media between the MTP and the held device with a=sendrecv. On call resumption, the MTP is removed from the media path. This feature addresses the mid-call Delayed Offer INVITE issue for audio direction, but it cannot resolve the issue of a device responding with its last used codec rather than the full list of all supported codecs. This issue can be problematic in cases where a codec change is required on re-establishment of media, such as placing a G.729 call on hold and connecting it to a music on hold source where G711 is preferred.

Require SDP Inactive in Mid-Call Media Exchange

SIP allows mid-call updates to codecs and connection information, such as IP addresses and UDP port numbers, without disconnecting the media connection. Some third-party devices cannot accept media changes using this method, and they require the media path to be closed gracefully and reopened to make media changes. If this feature is enabled, during mid-call codec or connection updates Cisco Unified CM sends an INVITE a=inactive SDP message to the endpoint to break the media exchange.

Note For SIP trunks enabled for Early Offer, this parameter will be overridden by the **Send send-receive SDP in mid-call INVITE** parameter.

Disable Early Media on 180

By default, Cisco Unified CM signals the registered calling phone to play local ringback if SDP is not received in a 180 Ringing or 183 Session Progress Response.

If SDP is included in the 180 or 183 Response, instead of playing ringback locally, Cisco Unified CM connects media, and the calling phone plays whatever the called device is sending in its media stream (such as ringback or busy signal).

If ringback is not received, the device to which you are connecting might be including SDP in the 180 response, but it is not sending any media before the 200 OK response. In this case, check this check box to play local ringback on the calling phone and connect the media upon receipt of the 200 OK response

Redirect by Application

When enabled, the Redirect by Application feature allows Unified CM to:

- Apply a specific calling search space to redirected contacts that are received in the 3xx response.
- Apply digit analysis to the redirected contacts to make sure that the call gets routed correctly.
- Prevent a DOS attack by limiting the number of redirection (recursive redirection) requests
- Allow other features to be invoked while the redirection is taking place.

If the Redirect by Application check box is unchecked, outbound SIP trunk calls can be redirected to a restricted phone number (such as an international number) because redirection is handled and routed at the SIP stack level without intervention of Unified CM digit analysis and class of service.

Re-Route Incoming Request to New Trunk Based on

Inbound SIP trunk calls to Unified CM will be accepted only if the source IP address and port number of the calling device match the destination IP address and port number of a configured SIP trunk. Once the call has been accepted, it can then optionally be re-routed to another Unified CM SIP trunk based on information contained within the received SIP message header.

By default, calls are never re-routed after being matched to a SIP trunk based on IP address and port number.

Optionally, incoming Requests can be re-routed to a new trunk based on the received:

- Contact header

  The call is re-routed to another SIP trunk based on the IP address and port number received in the contact header. This feature is typically used to re-route calls from a SIP Proxy to a Unified CM SIP trunk assigned to a specific end user or system.

- Call-Info Header with purpose=x-cisco-origIP

  This option is used to match inbound calls from Cisco Unified Customer Voice Portal (CVP) to a specific trunk based on the IP address and port number contained in the Call-Info header parameter purpose=x-cisco-origIP. This feature is typically used to map calls from Unified CVP to Unified CM trunks for call admission control.

Overwriting Caller ID Number and Name in Outbound Trunk Calls

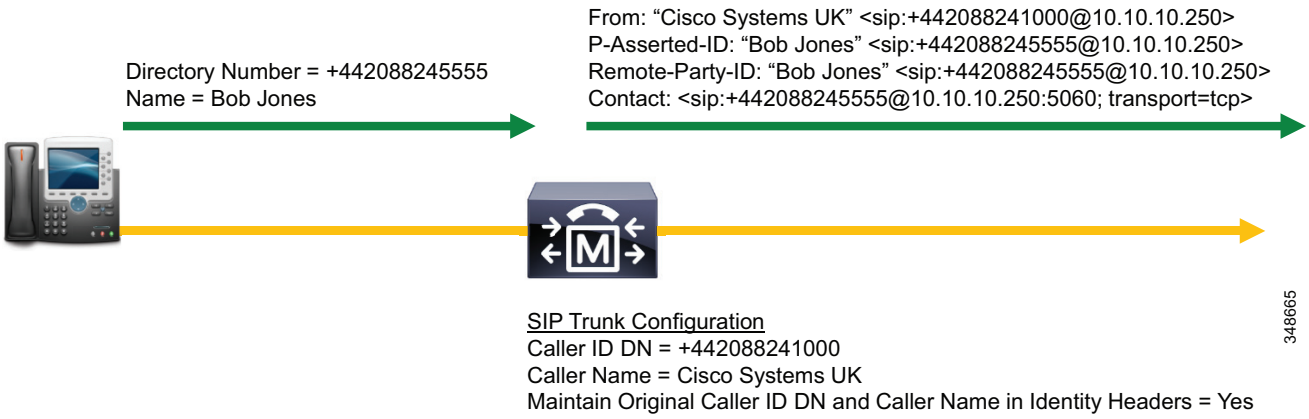
This feature can be useful if, for example, you wish to send a company switchboard number and company name instead of the caller's number and name in the SIP messages of calls sent over the SIP trunk. (See Figure 6-30.)This feature can be applied at the device level (for branch offices using a centralized SIP trunk) or at the trunk level.

At the device level, use the Caller ID DN and Caller Name fields of the Incoming Requests FROM URI Setting section on the SIP Profile associated with the device.

At the trunk level, use the Caller ID DN and Caller Name fields of the Outbound Calls - Caller Information section of the trunk configuration page.

By default the Caller ID DN and Caller Name sent in the From header, Contact header, and P-Asserted-Identity and Remote-Party-ID headers are modified in outbound SIP trunk calls. If you wish to keep the original Caller ID in the P-Asserted-Identity and Remote-Party-ID headers, check the **Maintain Original Caller ID DN and Caller Name in Identity Headers** check box on the trunk configuration page. Checking this check box allows the originator of the call to be traced.

Figure 6-30 Overwriting Caller ID Number and Caller Name on Outbound SIP Trunk Calls



SIP Trunk Message Normalization and Transparency

Normalization and transparency provide powerful script-based functionality for SIP trunks that can be used transparently to forward and/or modify SIP messages and message body (SDP) contents as they traverse Unified CM. Normalization and transparency scripts are designed to address SIP interoperability issues, allowing Unified CM to interoperate with SIP-based third-party PBXs, applications, and IP PSTN services.

SIP Trunk Normalization

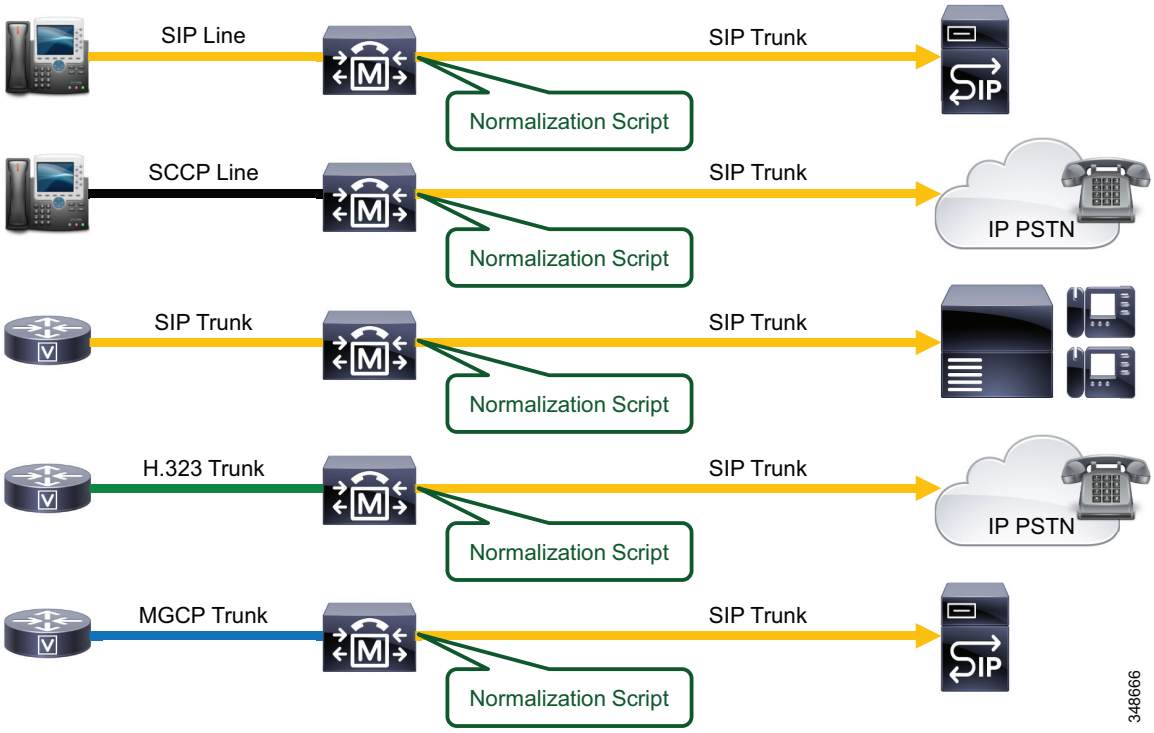
Normalization allows incoming and outgoing SIP messages to be modified on their way through Unified CM. For example, Unified CM supports the Diversion header for carrying redirecting number information. Some SIP devices connected to Unified CM use the History-Info header for this purpose. An inbound normalization script can be used to transform the History-Info headers into Diversion headers so that Unified CM recognizes the redirecting information. Likewise, an outbound normalization script can be used to transform Diversion headers into History-Info headers so that the external SIP device will recognize the redirecting information.

The normalization script is associated with a SIP trunk or a SIP line. The scripts manifest themselves as a set of message handlers that operate on inbound and outbound SIP messages. For normalization, the script manipulates almost every aspect of a SIP message, including:

- The request URI
- The response code and phrase
- SIP headers
- SIP parameters
- Content bodies
- SDP

Normalization applies to all calls that traverse a SIP trunk with an associated script, regardless of what protocol is being used for the other endpoint involved in the call. For example, a SIP trunk normalization script can operate on a call from a SIP line device to a SIP trunk, from an SCCP device to a SIP trunk, from MGCP trunk to SIP trunk, from H.323 trunk to SIP trunk, and so forth. (See Figure 6-31.)

Figure 6-31 SIP Trunk Normalization



SIP Trunk Transparency

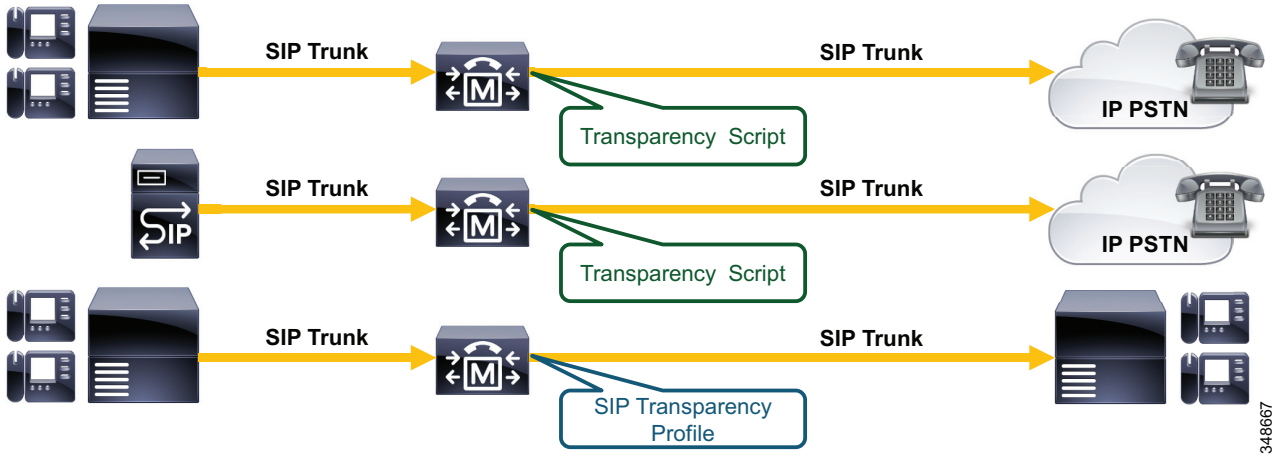
Transparency Lua scripts allow Unified CM to pass SIP headers, parameters, and message body contents from one SIP trunk call leg to another, even if Unified CM does not understand or support the parts of the message that are being passed through. Transparency (or transparent pass-through) is applicable only for SIP-to-SIP calls through Unified CM. (See Figure 6-32.)

The transparency script is associated with a SIP trunk or a SIP line. The scripts manifest themselves as a set of message handlers that operate on inbound and outbound SIP messages. For transparency, the script passes through almost any information in a SIP message, including:

- SIP headers
- SIP parameters
- Content bodies

SIP Profiles also support SDP Transparency Profiles, which can be used to pass either all unknown SDP parameters (default) or selected SDP parameters that are not natively supported by Unified CM, from one SIP trunk (or SIP endpoint) to another without using Lua transparency scripts.

Figure 6-32 SIP Trunk Transparency



Normalization and transparency scripts use Lua, a powerful, fast and lightweight, embeddable scripting language to modify SIP messages and SDP body content on SIP trunks. (For more information on Lua, refer to the documentation available at https://lua-users.org/wiki/LuaOrgGuide.)

For more information on SIP trunk normalization and transparency scripts, refer to the latest version of the *Developer Guide for SIP Transparency and Normalization*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps556/products_programming_reference_guides_list.html

The developer guide describes the scripting environment and APIs used to manipulate and pass through SIP message information.

For more information on script management, refer to the latest version of the *Cisco Unified Communications Manager Administration Guide*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Pre-Loaded Unified CM Normalization and Transparency Scripts

A number of normalization and transparency scripts are pre-loaded into Unified CM, and the following scripts are a representative sample of them:

- Refer-passthrough script — This script allows Unified CM to be removed from the call signaling path when a blind transfer (using an in-dialog REFER) is invoked between two SIP trunks.
- ContactHeader script — This script removes the audio and video attributes from the contact header in an inbound Delayed Offer mid-call re-invite.
- HCS-PCV-PAI-passthrough script — This script is used for integration with IP Multimedia Subsystem (IMS) networks, and it passes through or adds the P-Charging-Vector header in INVITE, UPDATE, and 200 OK messages.
- Diversion-Counter script — This script provides the capability to adjust the diversion counter for various Call Forward scenarios.
- VCS-interop script — This script provides interoperability for endpoints registered to the Cisco TelePresence Video Communication Server (VCS).

IP PSTN and IP Trunks to Service Provider Networks

Service providers are increasing their offerings of non-TDM PSTN connections to enterprise customers. Apart from the key benefit of the cost savings from deploying non-TDM interfaces, these IP-based PSTN connections can also offer additional voice features compared to traditional PSTN interfaces.

SIP services dominate today's available offerings, and although earlier H.323 services were available in select geographies, they are being phased out. This is mainly due to the increasing popularity of SIP as the protocol of choice by unified communications vendors and within the enterprise.

When connecting to a service provider's IP PSTN network, Cisco strongly recommends the use of the Cisco Unified Border Element as an enterprise edge Session Border Controller to provide a controlled demarcation and security point between your enterprise and the service provider's network.

Cisco Unified Border Element

Cisco Unified Border Element is a Session Border Controller that provides the following features and services:

- Address and port translations (privacy and Level 7 topology hiding)
- Conversion from SIP Delayed Offer to Early Offer
- Protocol interworking (H.323 and SIP) and normalization
- Media interworking (DTMF translation, fax, transcoding, transrating, volume and gain control)
- Call admission control (based on total calls, CPU, memory, call arrival spike detection, or maximum calls per destination)
- Security (including RTP to SRTP interworking, SIP malformed packet detection, non-dialog RTP packet drops, SIP listening port configuration, digest authentication, simultaneous call limits, call rate limits, toll fraud protection, and a number of signaling and media encryption options)
- PPI/PAI/Privacy and RPID – Identity Header Interworking with service providers
- QoS and bandwidth management (QoS marking using ToS, DSCP, and bandwidth enforcement using RSVP and codec filtering)
- Simultaneous connectivity to SIP trunks from multiple service providers
- High availability with in-box or box-to-box failover options (platform dependant)
- URI routing use GDPR route-strings to match dial peers
- Domain-based routing
- Multicast music on hold to unicast music on hold
- Voice and video media forking
- Enterprise Phone Proxy – VPN-Less IP Phone registration to Unified CM through Cisco Unified Border Element
- Billing statistics and CDR collection

The Cisco Unified Border Element is a licensed Cisco IOS application available on a wide range of Cisco router and gateway platforms. Depending on your choice of hardware platform, the Cisco Unified Border Element can provide session scalability from 4 to 16,000 concurrent voice calls with in-box or box-to-box failover options.

For more information on the Cisco Unified Border Element, refer to the documentation at

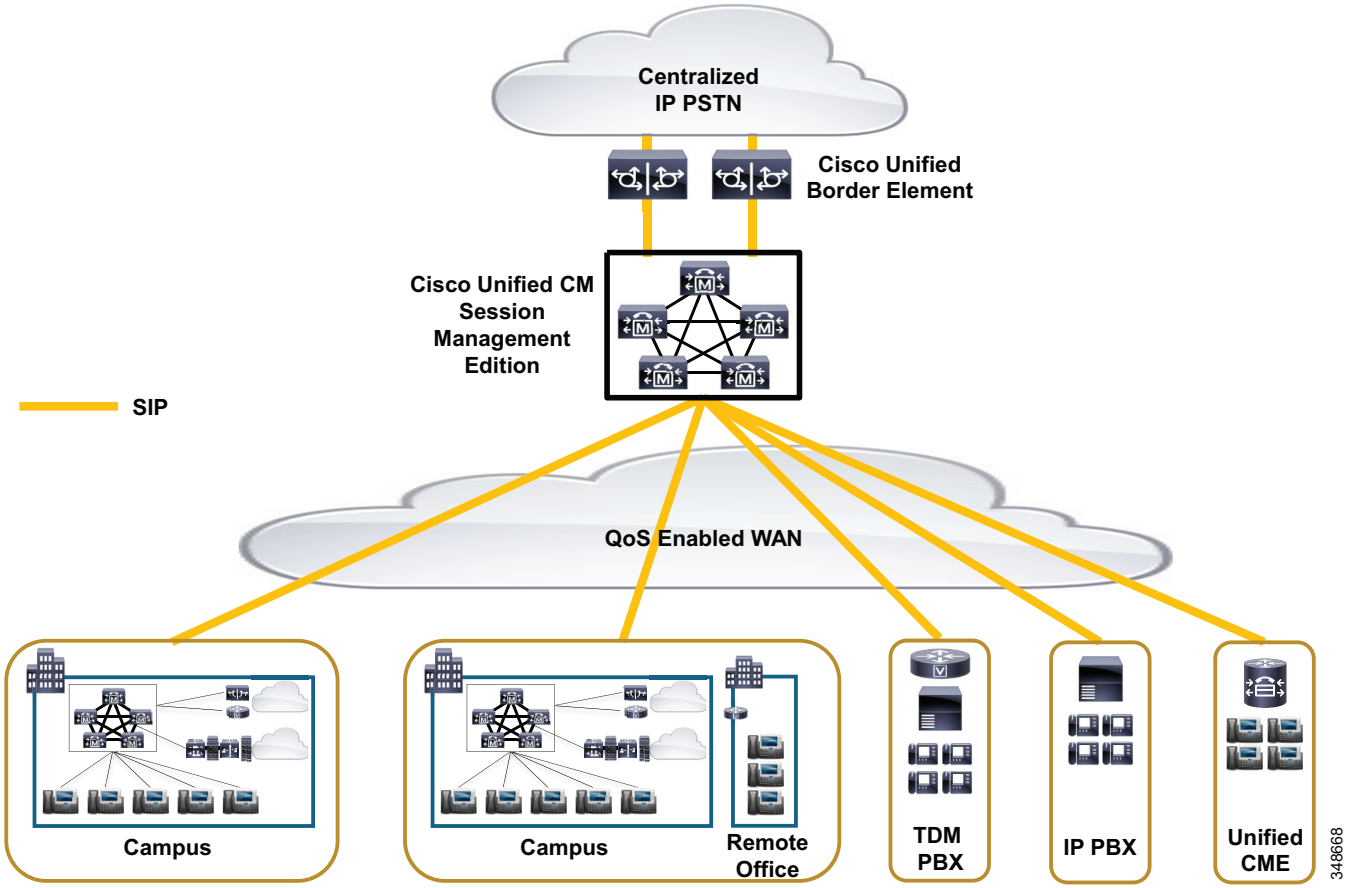
https://www.cisco.com/go/cube

IP-PSTN Trunk Connection Models

Trunks may be connected to IP PSTN service providers in several different ways, depending on the desired architecture. The two most common architectures for this connectivity are centralized trunks and distributed trunks.

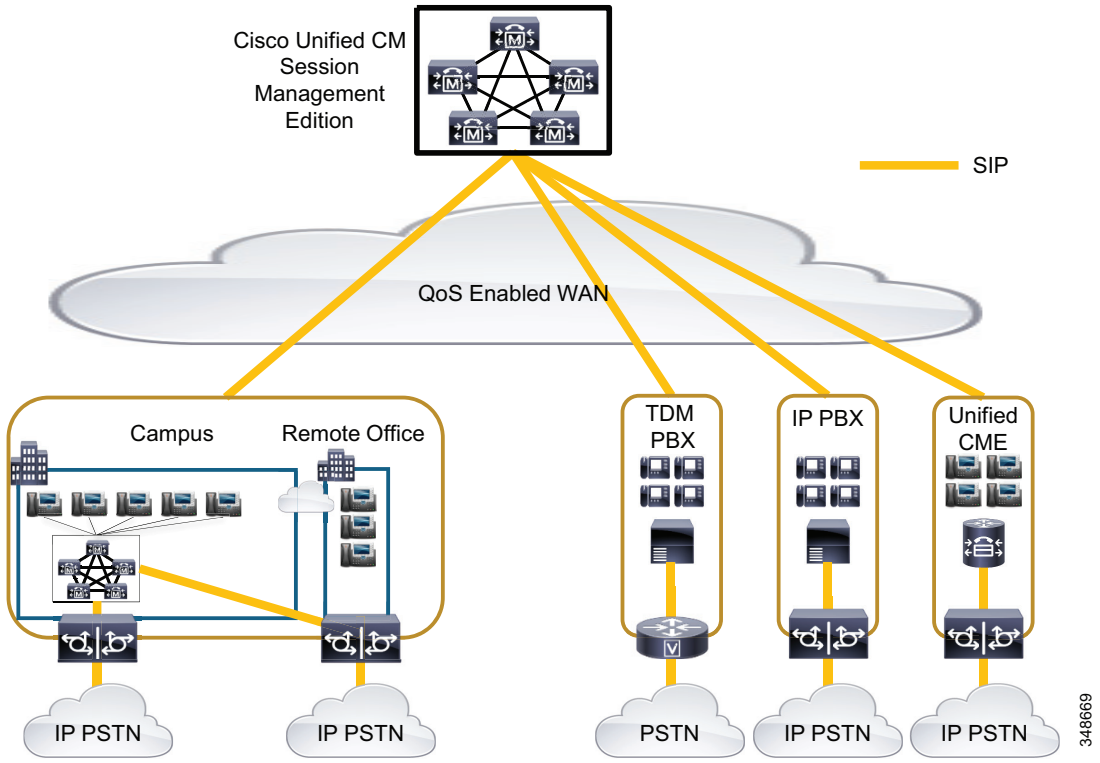
Centralized trunks connect the enterprise network to the service provider through one logical connection (although there may be more than one physical connection for redundancy) by means of a Session Border Controller (SBC) such as the Cisco Unified Border Element. (See Figure 6-33.) All calls to and from the IP PSTN use this set of trunks. For all sites remote from the centralized IP PSTN connection, the media and signaling for PSTN calls must traverse the enterprise IP WAN.

Figure 6-33 Centralized or Aggregated SIP Trunk Model



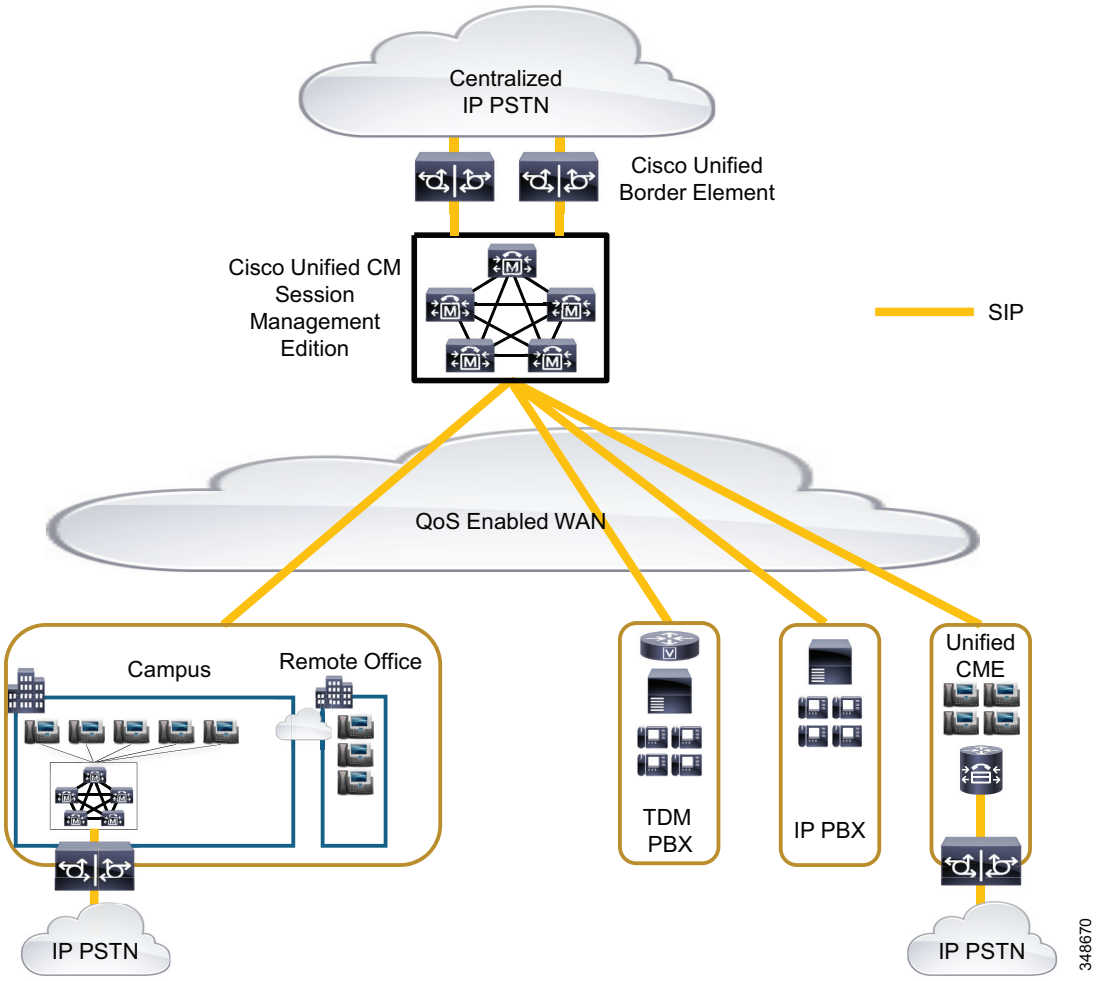
Distributed trunks connect to the service provider through several geographically distributed logical connections. (See Figure 6-34.) Each branch of an enterprise may have its own local trunk to the service provider. With distributed trunks in each branch site, media no longer needs to traverse the enterprise WAN, but flows to the service provider interface through a local SBC.

Figure 6-34 Distributed SIP Trunk Model



Each connectivity model has its own advantages and disadvantages. Centralized trunks are generally easier to deploy in terms of both physical equipment and configuration complexity, but media and signaling must traverse the enterprise to reach the PSTN, therefore requiring high availability in the enterprise WAN. Distributed trunks have the advantage of local hand-off of media and better number portability from local providers. As illustrated in Figure 6-35, a hybrid connectivity model that groups some of the branches together for connectivity, or that provides trunks from each Unified CM cluster of a multi-cluster deployment, captures the advantages of both forms of deployment models.

Figure 6-35 Hybrid SIP Trunk Model with Regional Aggregation



IP PSTN Trunks and Emergency Services

IP trunks might be unable to deliver emergency 911 calls or, like centralized PSTN trunks, might be unable to deliver such calls to the appropriate Public Safety Answering Point (PSAP) for the caller's location. Customers must investigate carefully the capabilities of the IP trunk service provider to deliver emergency 911 calls and caller locations to the appropriate PSAP. Cisco Emergency Responder may be used to provide the location-specific calling party number to the IP trunk service provider for emergency 911 calls.

Centralized IP or PSTN trunks might also temporarily become unavailable for emergency 911 calls from remote locations due to WAN congestion or failure. For this reason, remote locations should always have local gateways to the PSTN that are capable of delivering emergency 911 calls. For more information, see the chapter on Emergency Services, page 15-1.

Media Resources

Revised: March 1, 2018

A media resource is a software-based or hardware-based entity that performs media processing functions on the data streams to which it is connected. Media processing functions include mixing multiple streams to create one output stream (conferencing), passing the stream from one connection to another (media termination point), converting the data stream from one compression type to another (transcoding), streaming music to callers on hold (music on hold), echo cancellation, signaling, voice termination from a TDM circuit (coding/decoding), packetization of a stream, streaming audio (annunciation), and so forth. The software-based resources are provided by the Cisco Unified Communications Manager (Unified CM) IP Voice Media Streaming Service (IP VMS). Digital signal processor (DSP) cards provide both software and hardware based resources.

This chapter explains the overall Cisco Unified CM media resources architecture and Cisco IP Voice Media Streaming Application service, and it focuses on the following media resources:

- Voice Termination, page 7-4
- Transcoding, page 7-5
- Media Termination Point (MTP), page 7-7
- Trusted Relay Point, page 7-15
- Annunciator, page 7-15
- Cisco RSVP Agent, page 7-17
- Music on Hold, page 7-17

Use this chapter to gain an understanding of the function and capabilities of each media resource type available on Unified CM and to determine which resource are required for your deployment. For information on conferencing resources, refer to the chapter on Cisco Rich Media Conferencing, page 11-1.

For proper DSP sizing of Cisco Integrated Service Router (ISR) gateways, Cisco employees and partners with a valid login account can use the following tools:

- Cisco Collaboration Sizing Tool, available at https://cucst.cloudapps.cisco.com/landing
- DSP Calculator, available at https://www.cisco.com/c/en/us/applications/dsp-calc.html

Media Resources Architecture

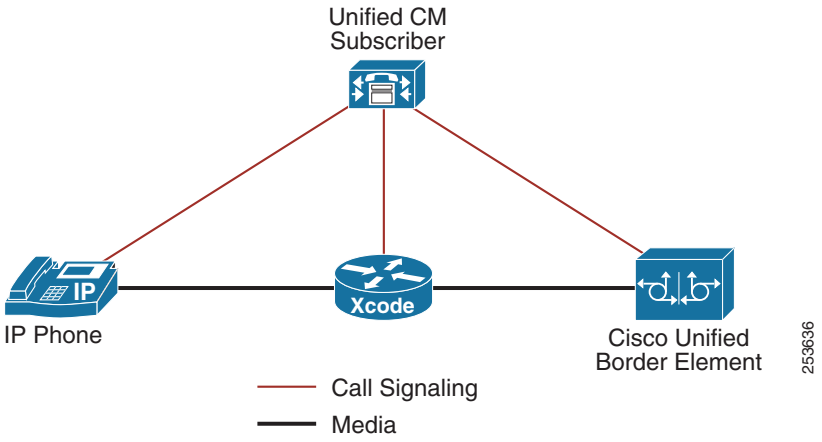
To properly design the media resource allocation strategy for an enterprise, it is critical to understand the Cisco Unified CM architecture for the various media resource components. The following sections highlight the important characteristics of media resource design with Unified CM.

Media Resource Manager

The Media Resource Manager (MRM), a software component in the Unified CM, determines whether a media resource needs to be allocated and inserted in the media path. This media resource may be provided by the Unified CM IP Voice Media Streaming Application service or by digital signal processor (DSP) cards. When the MRM decides and identifies the type of the media resource, it searches through the available resources according to the configuration settings of the media resource group list (MRGL) and media resource groups (MRGs) associated with the devices in question. MRGLs and MRGs are constructs that hold related groups of media resources together for allocation purposes and are described in detail in the section on Media Resource Groups and Lists, page 7-34.

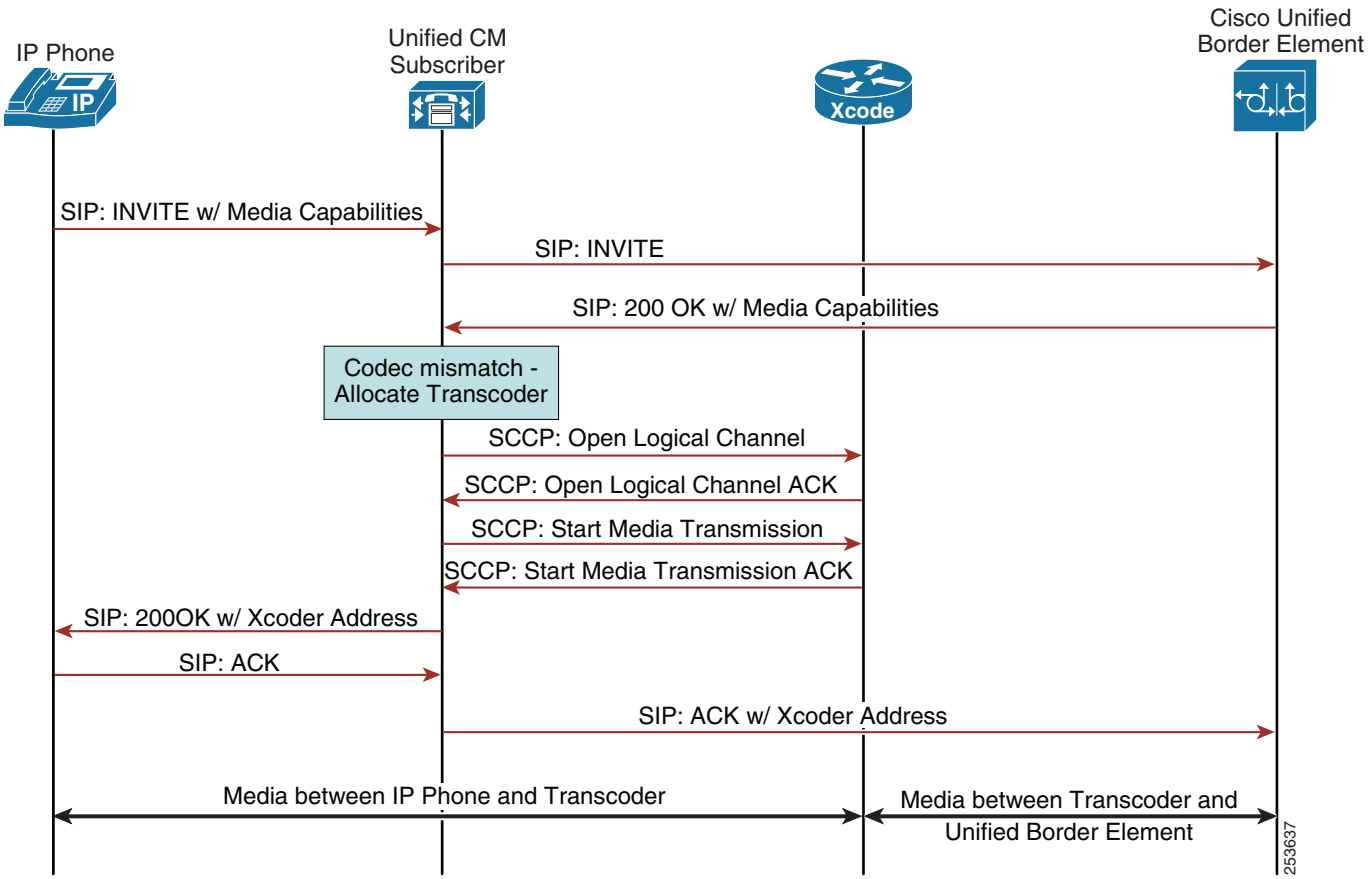
Figure 7-1 shows how a media resource such as a transcoder may be placed in the media path between an IP phone and a Cisco Unified Border Element when a common codec between the two is not available.

Figure 7-1 Use of a Transcoder Where a Common Codec Is Not Available



Unified CM communicates with media resources using Skinny Client Control Protocol (SCCP). This messaging is independent of the protocol that might be in use between Unified CM and the communicating entities. Figure 7-2 shows an example of the message flow, but it does not show all of the SCCP or SIP messages exchanged between the entities.

Figure 7-2 Message Flow Between Components



Cisco IP Voice Media Streaming Application

The Cisco IP Voice Media Streaming Application provides the following software-based media resources:

- Conference bridge
- Music on Hold (MoH)
- Annunciator
- Media termination point (MTP)
- Interactive Voice Response (IVR)

When the IP Voice Media Streaming Application is activated, one of each of the above resources is automatically configured. Conferencing, annunciator, IVR, and MTP services can be disabled if required. If these resources are not needed, Cisco recommends that you disable them by modifying the appropriate service parameter in the Unified CM configuration. The service parameters have default settings for the maximum number of connections that each media device can handle. For details on how to modify the service parameters, refer to the appropriate version of the *Cisco Unified Communications Manager Administration Guide*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Give careful consideration to situations that require multiple resources and to the load they place on the IP Voice Media Streaming Application. The media resources can reside on the same server as Unified CM or on a dedicated server not running the Unified CM call processing service. If your deployment requires more than the default number of any resource, Cisco recommends that you configure that resource to run on its own dedicated server. Cisco strongly recommends that you do not activate the Cisco IP Voice Streaming Media Application on a Cisco Unified CM node that has a high call processing load because it can adversely affect the performance of Cisco Unified CM. If heavy use of media resources is expected within a deployment, Cisco recommends deploying dedicated Unified CM media resource nodes (non-publisher nodes that do not perform call processing within the cluster) or relying on hardware-based media resources. Software-based media resources on Unified CM nodes are intended for small deployments or deployments where need for media resources is limited.

Voice Termination

Voice termination applies to a call that has two call legs, one leg on a time-division multiplexing (TDM) interface and the second leg on a Voice over IP (VoIP) connection. The TDM leg must be terminated by hardware that performs encoding/decoding and packetization of the stream. This termination function is performed by a digital signal processor (DSP) resource residing in the same hardware module, blade, or platform.

All DSP hardware on Cisco TDM gateways is capable of terminating voice streams, and certain hardware is also capable of performing other media resource functions such as conferencing or transcoding (see Transcoding, page 7-5 and Transcoding, page 7-5). The DSP hardware has either fixed DSP resources that cannot be upgraded or changed, or modular DSP resources that can be upgraded.

The number of supported calls per DSP depends on the computational complexity of the codec used for a call and also on the complexity mode configured on the DSP. Cisco IOS enables you to configure a complexity mode on the hardware module. Hardware platforms such as the PVDM2, PVDM3, and PVDM4 DSPs support three complexity modes: medium, high and flex mode. Some of the other hardware platforms support only medium and high complexity modes.

Medium and High Complexity Mode

You can configure each DSP separately as either medium complexity, high complexity, or flex mode (PVDM3 DSPs and those based on C5510). The DSP treats all calls according to its configured complexity, regardless of the actual complexity of the codec of the call. A resource with configured complexity equal or higher than the actual complexity of the incoming call must be available, or the call will fail. For example, if a call requires a high-complexity codec but the DSP resource is configured for medium complexity mode, the call will fail. However, if a medium-complexity call is attempted on a DSP configured for high complexity mode, then the call will succeed and Cisco IOS will allocate a high-complexity mode resource.

Flex Mode

Flex mode, available on hardware platforms that use the C5510 chipset and on PVDM3 DSPs, eliminates the requirement to specify the codec complexity at configuration time. A DSP in flex mode accepts a call of any supported codec type, as long as it has available processing power.

For C5510-based DSPs, the overhead of each call is tracked dynamically via a calculation of processing power in millions of instructions per second (MIPS). Cisco IOS performs a MIPS calculation for each call received and subtracts MIPS credits from its budget whenever a new call is initiated. The number of MIPS consumed by a call depends on the codec of the call. The DSP will allow a new call as long as it has remaining MIPS credits greater than or equal to the MIPS required for the incoming call.

Similarly, PVDM3 DSP modules use a credit-based system. Each module is assigned a fixed number of "credits" that represent a measure of its capacity to process media streams. Each media operation, such as voice termination, transcoding, and so forth, is assigned a cost in terms of credits. As DSP resources are allocated for a media processing function, its cost value is subtracted from the available credits. A DSP module runs out of capacity when the available credits run out and are no longer sufficient for the requested operation. The credit allocation rules for PVDM3 DSPs are rather complex.

Flex mode has an advantage when calls of multiple codecs must be supported on the same hardware because flex mode can support more calls than when the DSPs are configured as medium or high complexity. However, flex mode does allow oversubscription of the resources, which introduces the risk of call failure if all resources are used. With flex mode it is possible to have fewer DSP resources than with physical TDM interfaces.

Compared to medium or high complexity mode, flex mode has the advantage of supporting the most G.711 calls per DSP. For example, a PVDM2-16 DSP can support 8 G.711 calls in medium complexity mode or 16 G.711 calls in flex mode.

Transcoding

A transcoder is a device that converts an input stream from one codec into an output stream that uses a different codec. Starting with Cisco IOS Release 15.0.1M, a transcoder also supports transrating, whereby it connects two streams that utilize the same codec but with a different packet size.

Transcoding from G.711 to any other codec is referred to as traditional transcoding. Transcoding between any two non-G.711 codecs is called universal transcoding and requires Universal Cisco IOS transcoders. Universal transcoding is supported starting with Cisco IOS Release 12.4.20T. Universal transcoding has a lower DSP density than traditional transcoding.

In a Unified CM system, the typical use of a transcoder is to convert between a G.711 voice stream and the low bit-rate compressed voice stream G.729a. The following cases determine when transcoder resources are needed:

- Single codec for the entire system

  A single codec is generally used in a single-site deployment that usually has no need for conserving bandwidth. When a single codec is configured for all calls in the system, then no transcoder resources are required. In this scenario, G.711 is the most common choice that is supported by all vendors.

- Multiple codecs in use in the system, with all endpoints capable of all codec types

  The most common reason for multiple codecs is to use G.711 for LAN calls to maximize the call quality and to use a low-bandwidth codec to maximize bandwidth efficiency for calls that traverse a WAN with limited bandwidth. Cisco recommends using G.729a as the low-bandwidth codec because it is supported on all Cisco Unified IP Phone models as well as most other Cisco Unified Communications devices, therefore it can eliminate the need for transcoding. Although Unified CM allows configuration of other low-bandwidth codecs between regions, some phone models do not support those codecs and therefore would require transcoders. They would require one transcoder for a call to a gateway and two transcoders if the call is to another IP phone. The use of transcoders is avoided if all devices support and are configured for both G.711 and G.729 because the devices will use the appropriate codec on a call-by-call basis.

- Multiple codecs in use in the system, and some endpoints support or are configured for G.711 only

  This condition exists when G.729a is used in the system but there are devices that do not support this codec, or a device with G.729a support may be configured to not use it. In this case, a transcoder is also required. Devices from some third-party vendors may not support G.729.

A transcoder is also capable of performing the same functionality as a media termination point (MTP). In cases where transcoder functionality and MTP functionality are both needed, a transcoder is allocated by the system. If MTP functionality is required, Unified CM will allocate either a transcoder or an MTP from the resource pool, and the choice of resource will be determined by the media resource groups, as described in the section on Media Resource Groups and Lists, page 7-34.

To finalize the design, it is necessary to know how many transcoders are needed and where they will be placed. For a multi-site deployment, Cisco recommends placing a transcoder local at each site where it might be required. If multiple codecs are needed, it is necessary to know how many endpoints do not support all codecs, where those endpoints are located, what other groups will be accessing those resources, how many maximum simultaneous calls these device must support, and where those resources are located in the network.

Audio Transcoding Resources

Digital signal processor (DSP) resources are required to perform transcoding. The DSP resources can be located in the voice modules or in the on-board Cisco Packet Voice/Fax Digital Signal Processor (PVDM2, PVDM3, or PVDM4) slots available on the Cisco Integrated Services Routers (ISRs).

For more information on the voice gateway and its supported voice modules, refer the information on Cisco Voice Modules and Interface Cards available at

https://www.cisco.com/c/en/us/products/interfaces-modules/voice-modules-interface-cards/index.html

Video Interoperability

Video interoperability is the audio and video support for point-to-point calls between Cisco TelePresence System (CTS) endpoints, other Cisco Unified Communications video endpoints, and third-party video endpoints. Prior to Cisco Unified CM 8.5, video interoperability between the various families of video endpoints was possible only with the insertion of a video component between endpoints, such as a video transcoder or a multipoint control unit (MCU).

Cisco Unified CM 8.5 and later releases not only offer native video interoperability between the various video endpoint family types, point-to-point, but also provide better video interoperability in general with H.264 codec negotiation in SIP and H.323 protocols and enable the endpoints to negotiate high definition (HD) resolutions when available. Video interoperability, however, is dependent on the endpoints to support the interoperation. For further information, refer to *Interoperability Between CTS Endpoints and Other Cisco Endpoints or Devices*, available at

https://www.cisco.com/en/US/docs/telepresence/interop/endpoint_interop.html

Media Termination Point (MTP)

A media termination point (MTP) is an entity that accepts two full-duplex media streams. It bridges the streams together and allows them to be set up and torn down independently. The streaming data received from the input stream on one connection is passed to the output stream on the other connection, and vice versa. MTPs have many possible uses, such as:

- Re-Packetization of a Stream, page 7-7
- DTMF Conversion, page 7-7
- Protocol-specific usage (bridging between IPv4 and IPv6 endpoints)
  - Calls over SIP Trunks, page 7-9
  - H.323 Supplementary Services, page 7-12
  - H.323 Outbound Fast Connect, page 7-12

Re-Packetization of a Stream

An MTP can be used to transcode G.711 a-law audio packets to G.711 mu-law packets and vice versa, or it can be used to bridge two connections that utilize different packetization periods (different sample sizes).

DTMF Conversion

DTMF tones are used during a call to signal to a far-end device for purposes of navigating a menu system, entering data, or other manipulation. They are processed differently than DTMF tones sent during a call setup as part of the call control. There are several methods for sending DTMF over IP, and two communicating endpoints might not support a common procedure. In these cases, Unified CM may dynamically insert an MTP in the media path to convert DTMF signals from one endpoint to the other. Unfortunately, this method does not scale because one MTP resource is required for each such call. The following sections help determine the optimum amount of MTP resources required, based on the combination of endpoints, trunks, and gateways in the system.

If Unified CM determines that an MTP needs to be inserted but no MTP resources are available, it uses the setting of the service parameter **Fail call if MTP allocation fails** to decide whether or not to allow the call to proceed.This service parameter is set to a default value of **False**. With this default configuration, an incoming call on a SIP Early Offer trunk would result in an outbound Delayed Offer.

DTMF Relay Between Endpoints

The following methods are used to relay DTMF from one endpoint to another.

Named Telephony Events (RFC 2833)

Named Telephony Events (NTEs) defined by RFC 2833 are a method of sending DTMF from one endpoint to another after the call media has been established. The tones are sent as packet data using the already established RTP stream and are distinguished from the audio by the RTP payload type field. For example, the audio of a call can be sent on a session with an RTP payload type that identifies it as G.711 data, and the DTMF packets are sent with an RTP payload type that identifies them as NTEs. The consumer of the stream utilizes the G.711 packets and the NTE packets separately.

Key Press Markup Language (RFC 4730)

The Key Press Markup Language (KPML) is defined in RFC 4730. Unlike NTEs, which is an in-band method of sending DTMF, KPML uses the signaling channel (out-of-band, or OOB) to send SIP messages containing the DTMF digits.

KPML procedures use a SIP SUBSCRIBE message to register for DTMF digits. The digits themselves are delivered in NOTIFY messages containing an XML encoded body.

Unsolicited Notify (UN)

Unsolicited Notify procedures are used primarily by Cisco IOS SIP Gateways to transport DTMF digits using SIP NOTIFY messages. Unlike KPML, these NOTIFY messages are unsolicited, and there is no prior registration to receive these messages using a SIP SUBSCRIBE message. But like KPML, Unsolicited Notify messages are out-of-band.

Also unlike KPML, which has an XML encoded body, the message body in these NOTIFY messages is a 10-character encoded digit, volume, and duration, describing the DTMF event.

H.245 Signal, H.245 Alphanumeric

H.245 is the media control protocol used in H.323 networks. In addition to its use in negotiating media characteristics, H.245 also provides a channel for DTMF transport. H.245 utilizes the signaling channel and, hence, provides an out-of-band (OOB) way to send DTMF digits. The Signal method carries more information about the DTMF event (such as its actual duration) than does Alphanumeric.

Cisco Proprietary RTP

This method sends DTMF digits in-band, that is, in the same stream as RTP packets. However, the DTMF packets are encoded differently than the media packets and use a different payload type. This method is not supported by Unified CM but is supported on Cisco IOS Gateways.

Skinny Client Control Protocol (SCCP)

The Skinny Client Control Protocol is used by Unified CM for controlling the various SCCP-based devices registered to it. SCCP defines out-of-band messages that transport DTMF digits between Unified CM and the controlled device.

DTMF Relay Between Endpoints in the Same Unified CM Cluster

The following rules apply to endpoints registered to Unified CM servers in the same cluster:

- Calls between two non-SIP endpoints do not require MTPs.

  All Cisco Unified Communications endpoints other than SIP send DTMF to Unified CM via various signaling paths, and Unified CM forwards the DTMF between dissimilar endpoints. For example, an IP phone may use SCCP messages to Unified CM to send DTMF, which then gets sent to an H.323 gateway via H.245 signaling events. Unified CM provides the DTMF forwarding between different signaling types.

- Calls between two Cisco SIP endpoints do not require MTPs.

  All Cisco SIP endpoints support NTE, so DTMF is sent directly between endpoints and no conversion is required.

- A combination of a SIP endpoint and a non-SIP endpoint might require MTPs.

  To determine the support for NTE in your devices, refer to the product documentation for those devices. Support of NTE is not limited to SIP and can be supported in devices with other call control protocols. Unified CM has the ability to allocate MTPs dynamically on a call-by-call basis, based on the capabilities of the pair of endpoints.

Calls over SIP Trunks

SIP trunk configuration is used to set up communication with a SIP User Agent such as another Cisco Unified CM cluster or a SIP gateway.

SIP negotiates media exchange via Session Description Protocol (SDP), where one side offers a set of capabilities to which the other side answers, thus converging on a set of media characteristics. SIP allows the initial offer to be sent either by the caller in the initial INVITE message (Early Offer) or, if the caller chooses not to, the called party can send the initial offer in the first reliable response (Delayed Offer).

By default, Unified CM SIP trunks send the INVITE without an initial offer (Delayed Offer). Unified CM has the following three configurable options to enable a SIP trunk to send the offer in the INVITE (Early Offer):

Media Termination Point Required

Enabling this option on the SIP trunk assigns an MTP for every outbound call. This option does not support codec pass-through mode, which imposes a single codec (G.711 or G.729) limitation over the SIP trunk, thus limiting media to voice calls only. With this option enabled, calls over the trunk uses MTPs assigned to the trunk rather than using calling device MTPs, which forces the media to follow the same signaling path.

Note Enabling the **Media Termination Point Required** option on the SIP trunk increases MTP usage because an MTP is assigned for every inbound and outbound call rather than on an as-needed basis.

Early Offer support for voice and video calls Mandatory (insert MTP if needed)

Enabling this Unified CM configuration option on the SIP Profile associated with the SIP trunk inserts an MTP only if the calling device cannot provide Unified CM with the media characteristics required to create the outbound Early Offer (for example, where an inbound call to Unified CM is received on a Delayed Offer SIP trunk or a Slow Start H.323 trunk and on calls from older SCCP-based phones such as Cisco Unified IP Phones 7940 or 7960 registered to Unified CM). Unified CM creates a super-set of the endpoint and MTP codec capabilities and applies the codec filtering based on the applicable region-pair settings. The outbound Offer SDP will use the IP address and port number of the MTP and voice codec supported by the calling phone.

When Unified CM receives an inbound call on an H.323 Slow Start or SIP Delayed Offer trunk, the media capabilities of the calling device are not available when the call is initiated. In this case, Unified CM must insert an MTP and use its IP address and UDP port number to advertise all supported audio codecs (after region-pair filtering) in the Offer SDP of the initial INVITE sent over the outbound SIP trunk. When the Answer SDP is received on the SIP trunk, if it contains a codec that the calling endpoint supports, no additional offer-answer transaction is needed. In case of codec mismatch, Unified CM can either insert a transcoder to address the mismatch or send a Re-INVITE or UPDATE to trigger media negotiation. Calls from H.323 Slow Start or SIP Delayed Offer trunks support audio only in the initial call setup, but they can be upgraded mid-call to support video and SRTP if the call media is renegotiated (for example, after Hold/Resume).

When you configure **Early Offer support for voice and video calls Mandatory (insert MTP if needed)** on the SIP Profile of a trunk, calls from older SCCP-based phones, SIP Delayed Offer trunks, and H.323 Slow Start trunks cause Unified CM to allocate an MTP, if an MTP or transcoder is not already allocated for that call for another reason. The MTP is used to generate an Offer SDP with a valid media port number and IP address. The MTP is allocated from the media resources that are associated with the calling device rather than from the media resources of the outbound SIP trunk. (This prevents

the media path from being anchored to the MTP of the outbound SIP trunk.) If the MTP cannot be allocated from the media resource group list (MRGL) of the calling device, the MTP allocation is attempted from the MRGL of the SIP trunk.

Note If no MTP resources are available, the call will proceed as a Delayed Offer call.

Unified CM does not need to insert an MTP to create an outbound Early Offer call over a SIP trunk if Unified CM receives the inbound call by any of the following means:

- On a SIP trunk using Early Offer
- On an H.323 trunk using Fast Start
- On an MGCP trunk
- From a SIP-based IP phone registered to Unified CM

Early Offer support for voice and video calls Best Effort (No MTP inserted)

If this Unified CM SIP profile configuration option is enabled, the SIP trunk will never use MTPs to create an Early Offer but will send either an Early Offer or a Delayed Offer, depending on the capabilities of the calling device.

Best Effort Early Offer SIP trunks send outbound calls as Early Offer (INVITE with SDP content) in the following situations:

- An inbound call to Unified CM or SME is received over a SIP trunk using Early Offer.
- An inbound call to Unified CM or SME is received over an H.323 trunk using Fast Start.
- An inbound call to Unified CM or SME is received over an MGCP trunk.
- A call is initiated from a SIP-based IP phone registered to Unified CM.
- A call is initiated from a newer model SCCP-based Cisco Unified IP Phone registered to Unified CM.

Best Effort Early Offer trunks send outbound calls as Delayed Offer (INVITE without SDP content) in the following situations:

- An inbound call to Unified CM or SME is received over a Delayed Offer SIP trunk.
- An inbound call to Unified CM or SME is received over an H.323 Slow Start trunk.
- A call is initiated from an older model SCCP-based IP phone registered to Unified CM.

Calls over a Best Effort Early Offer SIP trunk support voice, video, and encrypted media.

In general, Cisco recommends **Early Offer support for voice and video calls Best Effort (No MTP inserted)** for all Unified CM and Unified CM Session Management Edition SIP trunks.

For more information on this option, refer to the section on Best Effort Early Offer [Early Offer support for voice and video calls Best Effort (no MTP inserted)], page 6-22.

SIP Trunk MTP Requirements

By default, the SIP trunk parameter **Media Termination Point Required** and the SIP Profile parameter **Early Offer support for voice and video calls** are not selected.

Use the following steps to determine whether MTP resources are required for your SIP trunks.

1. Is the far-end SIP device defined by this SIP trunk capable of accepting an inbound call without a SIP Early Offer?

   If not, then on the SIP Profile associated with this trunk, select **Early Offer support for voice and video calls (insert MTP if needed)**. For outbound SIP trunk calls, an MTP will be inserted only if the calling device cannot provide Unified CM with the media characteristics required to create the Early Offer, or if DTMF conversion is needed.

   If yes, then select **Early Offer support for voice and video calls Best Effort (No MTP inserted)**, and use Step 2. to determine whether an MTP is inserted dynamically for DTMF conversion. Note that DTMF conversion can be performed by the MTP regardless of the codec in use.

2. Select a Trunk DTMF Signaling Method, which controls the behavior of DTMF selection on that trunk. Available MTPs will be allocated based on the requirements for matching DTMF methods for all calls.

   a. DTMF Signaling Method: No Preference

      In this mode, Unified CM attempts to minimize the usage of MTP by selecting the DTMF signaling method supported by the endpoint.

      If both devices support RFC 2833, then no MTP is required.

      If both devices support any out-of-band DTMF mechanism, then Unified CM will use KPML over the SIP trunk. The only case where MTP is required is when one of the endpoints supports out-of-band only and the other supports RFC 2833 only.

      If both devices support RFC 2833 and any out-of-band DTMF mechanism, then Unified CM negotiates both RFC 2833 and KPML but relies on RFC 2833 to receive the digits.

      Note Unified CM does not negotiate to Unsolicited Notify if the SIP trunk is set to No Preference for its DTMF configuration. For example, if the far-end SIP gateway or Cisco Unified Border Element is configured for **dtmf-relay sip-notify** and the Unified CM SIP trunk to the Cisco Unified Border Element is set to No Preference, then DTMF will not work. The recommendation in this case is to set the Unified CM SIP trunk to OOB and RFC 2833 for the DTMF configuration, which will allow Unified CM to negotiate the Unsolicited Notify DTMF method with the SIP gateway or Cisco Unified Border Element.

   b. DTMF Signaling Method: RFC 2833

      By placing a restriction on the DTMF signaling method across the trunk, Unified CM is forced to allocate an MTP if any one or both the endpoints do not support RFC 2833. In this configuration, the only time an MTP will not be allocated is when both endpoints support RFC 2833.

   c. DTMF Signaling Method: OOB and RFC 2833

      In this mode, the SIP trunk signals both KPML and RFC 2833 DTMF methods across the trunk, and it is the most intensive MTP usage mode. The only cases where MTP resources will not be required is when both endpoints support RFC 2833 and any OOB DTMF method (KPML or SCCP).

Note Cisco Unified IP Phones play DTMF to the end user when DTMF is received via SCCP, but they do not play tones received by RFC 2833. However, there is no requirement to send DTMF to another end user. It is necessary only to consider the endpoints that originate calls combined with endpoints that might need DTMF, such as PSTN gateways, application servers, and so forth.

DTMF Relay on SIP Gateways and Cisco Unified Border Element

Cisco SIP Gateways support KPML, NTE, or Unsolicited Notify as the DTMF mechanism, depending on the configuration. Because there may be a mix of endpoints in the system, multiple methods may be configured on the gateway simultaneously in order to minimize MTP requirements.

On Cisco SIP Gateways, configure both **sip-kpml** and **rtp-nte** as DTMF relay methods under SIP dial peers. This configuration will enable DTMF exchange with all types of endpoints, including those that support only NTE and those that support only OOB methods, without the need for MTP resources. With this configuration, the gateway will negotiate both NTE and KPML with Unified CM. If NTE is not supported by the Unified CM endpoint, then KPML will be used for DTMF exchange. If both methods are negotiated successfully, the gateway will rely on NTE to receive digits and will not subscribe to KPML.

Cisco SIP gateways also have the ability to use proprietary Unsolicited Notify (UN) method for DTMF. The UN method sends a SIP Notify message with a body that contains text describing the DTMF tone. This method is also supported on Unified CM and may be used if **sip-kpml** is not available. Configure **sip-notify** as the DTMF relay method. Note that this method is Cisco proprietary.

SIP gateways that support only NTE require MTP resources to be allocated when communicating with endpoints that do not support NTE.

H.323 Trunks and Gateways

For the H.323 gateways and trunks there are three reasons for invoking an MTP:

- H.323 Supplementary Services, page 7-12
- H.323 Outbound Fast Connect, page 7-12
- DTMF Conversion, page 7-13

H.323 Supplementary Services

MTPs can be used to extend supplementary services to H.323 endpoints that do not support the H.323v2 OpenLogicalChannel and CloseLogicalChannel request features of the Empty Capabilities Set (ECS). This requirement occurs infrequently. All Cisco H.323 endpoints support ECS, and most third-party endpoints have support as well. When needed, an MTP is allocated and connected into a call on behalf of an H.323 endpoint. When an MTP is required on an H.323 call and none is available, the call will proceed but will not be able to invoke supplementary services.

H.323 Outbound Fast Connect

H.323 defines a procedure called Fast Connect, which reduces the number of packets exchanged during a call setup, thereby reducing the amount of time for media to be established. This procedure uses Fast Start elements for control channel signaling, and it is useful when two devices that are utilizing H.323 have high network latency between them because the time to establish media depends on that

latency. Unified CM distinguishes between inbound and outbound Fast Start based on the direction of the call setup, and the distinction is important because the MTP requirements are not equal. For inbound Fast Start, no MTP is required. Outbound calls on an H.323 trunk do require an MTP when Fast Start is enabled. Frequently, it is only inbound calls that are problematic, and it is possible to use inbound Fast Start to solve the issue without also enabling outbound Fast Start.

DTMF Conversion

An H.323 trunk supports the signaling of DTMF by means of H.245 out-of-band methods. H.323 intercluster trunks also support DTMF by means of NTE. There are no DTMF configuration options for H.323 trunks; Unified CM dynamically chooses the DTMF transport method.

The following scenarios can occur when two endpoints on different clusters are connected with an H.323 trunk:

- When both endpoints are SIP, then NTE is used. No MTP is required for DTMF.
- When one endpoint is SIP and supports both KPML and NTE, but the other endpoint is not SIP, then DTMF is sent as KPML from the SIP endpoint to Unified CM, and H.245 is used on the trunk. No MTP is required for DTMF.
- If one endpoint is SIP and supports only NTE but the other is not SIP, then H.245 is used on the trunk. An available MTP is allocated for the call. The MTP will be allocated on the Unified CM cluster where the SIP endpoint is located.

For example: A Cisco Unified IP Phone 7970 using SIP to communicate with a Cisco Unified IP Phone 7970 running SCCP, will use NTE when connected via a SIP trunk but will use OOB methods when communicating over an H.323 trunk (with the trunk using the H.245 method).

When a call is inbound from one H.323 trunk and is routed to another H.323 trunk, NTE will be used for DTMF when both endpoints are SIP. H.245 will be used if either endpoint is not SIP. An MTP will be allocated if one side is a SIP endpoint that supports only NTE and the other side is non-SIP.

DTMF Relay on H.323 Gateways and Cisco Unified Border Element

H.323 gateways support DTMF relay via H.245 Alphanumeric, H.245 Signal, NTE, and audio in the media stream. The NTE option must not be used because it is not supported on Unified CM for H.323 gateways at this time. The preferred option is H.245 Signal. MTPs are required for establishing calls to an H.323 gateway if the other endpoint does not have signaling capability in common with Unified CM. For example, a Cisco Unified IP Phone 7960 running the SIP stack supports only NTEs, so an MTP is needed with an H.323 gateway.

CTI Route Points

A CTI Route Point uses CTI events to communicate with CTI applications. For DTMF purposes, the CTI Route Point can be considered as an endpoint that supports all OOB methods and does not support RFC 2833. For such endpoints, the only instance where an MTP will be required for DTMF conversion would be when it is communicating with another endpoint that supports only RFC 2833.

CTI Route Points that have first-party control of a phone call will participate in the media stream of the call and require an MTP to be inserted. When the CTI has third-party control of a call so that the media passes through a device that is controlled by the CTI, then the requirement for an MTP is dependent on the capabilities of the controlled device.

Example 7-1 Call Flow that Requires an MTP for NTE Conversion

Assume the example system has CTI route points with first-party control (the CTI port terminates the media), which integrate to a system that uses DTMF to navigate an IVR menu. If all phones in the system are running SCCP, then no MTP is required. In this case Unified CM controls the CTI port and receives DTMF from the IP phones via SCCP. Unified CM provides DTMF conversion.

However, if there are phones running a SIP stack (that support only NTE and not KPML), an MTP is required. NTEs are part of the media stream; therefore Unified CM does not receive them. An MTP is invoked into the media stream and has one call leg that uses SCCP, and the second call leg uses NTEs. The MTP is under SCCP control by Unified CM and performs the NTE-to-SCCP conversion. Note that the newer phones that do support KPML will not need an MTP.

MTP Usage with a Conference Bridge

MTPs are utilized in a conference call when one or more participant devices in the conference use RFC 2833. When the conference feature is invoked, Unified CM allocates MTP resources for every conference participant device in the call that supports only RFC 2833. This is regardless of the DTMF capabilities of the conference bridge used.

MTP Resources

The following types of devices are available for use as an MTP:

Software MTP (Cisco IP Voice Media Streaming Application)

A software MTP is a device that is implemented by enabling the Cisco IP Voice Media Streaming Application on a Unified CM server. When the installed application is configured as an MTP application, it registers with a Unified CM node and informs Unified CM of how many MTP resources it supports. A software MTP device supports only G.711 streams or passthrough mode in a codec. The IP Voice Media Streaming Application is a resource that may also be used for several functions, and the design guidance must consider all functions together (see Cisco IP Voice Media Streaming Application, page 7-3).

Software MTP (Based on Cisco IOS)

- The capability to provide a software-based MTP on the router is available beginning with Cisco IOS Release 12.3(11)T for the Cisco 3800 Series Routers; Release 15.0(1)M for the Cisco 2900 Series and 3900 Series Routers; Release IOS-XE for ASR1002, 1004, and 1006 Routers; Release IOS-XE 3.2 for ASR1001 Routers; and Release 12.3(8)T4 for other router models.
- This MTP allows configuration of any of the following codecs, but only one may be configured at a given time: G.711 mu-law and a-law, G.729a, G.729, G.729ab, G.729b, and passthrough. Some of these are not pertinent to a Unified CM implementation.
- Router configurations permit up to 1,000 individual streams, which support 500 transcoded sessions. This number of G.711 streams generates 10 Mbytes of traffic. The Cisco ISR G2s and ASR routers can support significantly higher numbers than this.

Hardware MTP

DSP resources located in the voice modules or in the on-board Cisco Packet Voice/Fax Digital Signal Processor (PVDM2, PVDM3, or PVDM4) slots for Cisco Integrated Services Routers (ISRs) can also be used as MTP resource.

For more information on supported sessions with each PVDM module, refer the section on Capacity Planning for Media Resources, page 7-30.

Note When Cisco IOS MTP resources are invoked by Unified CM for a call flow, a software session rather than a hardware DSP session is consumed unless the media legs of the call flow require transrating. Thus, for flows invoking an MTP, a DSP session is used only when transrating (conversion between media legs with the same codec but different packetization times) is required.

Trusted Relay Point

A Trusted Relay Point (TRP) is a device that can be inserted into a media stream to act as a control point for that stream. It may be used to provide further processing on that stream or as a method to ensure that the stream follows a specific desired path. There are two components to the TRP functionality, the logic utilized by Unified CM to invoke the TRP and the actual device that is invoked as the anchor point of the call. The TRP functionality can invoke an MTP device to act as that anchor point.

Unified CM provides a new configuration parameter for individual phone devices, which invokes a TRP for any call to or from that phone. The system utilizes the media resource pool mechanisms to manage the TRP resources. The media resource pool of that device must have an available device that will be invoked as a TRP.

See the chapter on Network Infrastructure, page 3-1, for an example of a use case for the TRP as a QoS enforcement mechanism, and see the chapter on Cisco Collaboration Security, page 4-1, for an example of utilizing the TRP as an anchor point for media streams in a redundant data center with firewall redundancy.

Annunciator

An annunciator is a software function of the Cisco IP Voice Media Streaming Application that provides the ability to stream spoken messages or various call progress tones from the system to a user. It uses SCCP messages to establish RTP streams, and it can send multiple one-way RTP streams to devices such as Cisco IP phones or gateways. For most SIP devices, the call progress tones are downloaded (pushed) to the device at registration so that they can be invoked as needed by SIP signaling messages from Unified CM. Some SIP devices such as intercluster SIP trunks may still use an annunciator for call-progress tones. An annunciator may be used for verbal messages for almost any device regardless of whether it is using SIP or SCCP.

In some installations, it might be a requirement to establish a two-way media connection with an annunciator. To enable this capability, set the Cisco Unified CM service parameter **Duplex Streaming Enabled** to **True**. This may be required for firewall transversal or possibly for SIP early-offer scenarios.

Tones and announcements are predefined by the system. The announcements support localization and may also be customized by replacing the appropriate .wav file. The annunciator is capable of supporting G.711 a-law and mu-law, G.729, and Cisco L16 Wideband codecs without any transcoding resources.

The following features require an annunciator resource:

- Cisco Multilevel Precedence Preemption (MLPP)

  This feature has streaming messages that it plays in response to the following call failure conditions.

  - Unable to preempt due to an existing higher-precedence call.
  - A precedence access limitation was reached.
  - The attempted precedence level was unauthorized.
  - The called number is not equipped for preemption or call waiting.

- Integration via SIP trunk

  SIP endpoints have the ability to generate and send tones in-band in the RTP stream. Because SCCP devices do not have this ability, an annunciator is used in conjunction with an MTP to generate or accept DTMF tones when integrating with a SIP endpoint. The following types of tones are supported:

  - Call progress tones (busy, alerting, reorder, and ringback)
  - DTMF tones

- Cisco IOS gateways and intercluster trunks

  These devices require support for call progress tone (ringback tone).

- System messages

  During the following call failure conditions, the system plays a streaming message to the end user:

  - A dialed number that the system cannot recognize
  - A call that is not routed due to a service disruption
  - A number that is busy and not configured for preemption or call waiting

- Conferencing

  During a conference call, the system plays a barge-in tone to announce that a participant has joined or left the bridge.

An annunciator is automatically created in the system when the Cisco IP Voice Media Streaming Application is activated on a server. If the Media Streaming Application is deactivated, then the annunciator is also deactivated. A single annunciator instance can service the entire Unified CM cluster if it meets the performance requirements (see Annunciator Performance, page 7-16); otherwise, you must configure additional annunciators for the cluster. Additional annunciators can be added by activating the Cisco IP Voice Media Streaming Application on other servers within the cluster.

The annunciator registers with a single Unified CM at a time, as defined by its device pool and CM Group. It will automatically fail over to a secondary Unified CM if a secondary is configured for the device pool. Any announcement that is playing at the time of an outage will not be maintained.

An annunciator is considered a media device, and it can be included in media resource groups (MRGs) to control which annunciator is selected for use by phones and gateways.

Annunciator Performance

By default, the annunciator is configured to support 48 simultaneous streams, which is the maximum recommended for an annunciator running on the same server (co-resident) with the Unified CM service. If the server has only 10 Mbps connectivity, lower the setting to 24 simultaneous streams.

For more information on supported annunciator sessions with each server platform, refer to the section on Media Resources, page 25-28, in the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Cisco RSVP Agent

In order to provide topology-aware call admission control, Unified CM invokes one or two RSVP Agents during the call setup to perform an RSVP reservation across the IP WAN. These agents are MTP or transcoder resources that have been configured to provide RSVP functionality. RSVP resources are treated the same way as regular MTPs or transcoders from the perspective of allocation of an MTP or transcoder resource by Unified CM.

The Cisco RSVP Agent feature was first introduced in Cisco IOS Release 12.4(6)T. For details on RSVP and Cisco RSVP Agents, refer to the chapter on Bandwidth Management, page 13-1.

Music on Hold

The Music on Hold (MoH) feature requires that each MoH server must be part of a Unified CM cluster and participate in the data replication schema. Specifically, the MoH server must share the following information with the Unified CM cluster through the database replication process:

- Audio sources - The number and identity of all configured MoH audio sources
- Multicast or unicast - The transport nature (multicast or unicast) configured for each of these sources
- Multicast address - The multicast base IP address of those sources configured to stream as multicast

To configure a MoH server, enable the Cisco IP Voice Media Streaming Application Service on one or more Unified CM nodes. An MoH server can be deployed along with Unified CM on the same server or in standalone mode.

Unicast and Multicast MoH

Unified CM supports unicast and multicast MoH transport mechanisms.

A unicast MoH stream is a point-to-point, one-way audio Real-Time Transport Protocol (RTP) stream from the MoH server to the endpoint requesting MoH. It uses a separate source stream for each user or connection. Thus, if twenty devices are on hold, then twenty streams are generated over the network between the server and these endpoint devices. Unicast MoH can be extremely useful in those networks where multicast is not enabled or where devices are not capable of multicast, thereby still allowing an administrator to take advantage of the MoH feature. However, these additional MoH streams can potentially have a negative effect on network throughput and bandwidth.

A multicast MoH stream is a point-to-multipoint, one-way audio RTP stream between the MoH server and the multicast group IP address. The endpoints requesting an MoH audio stream can join the multicast group as needed. This mode of MoH conserves system resources and bandwidth because it enables multiple users to use the same audio source stream to provide music on hold. For this reason, multicast is an extremely attractive transport mechanism for the deployment of a service such as MoH because it greatly reduces the CPU impact on the source device and also greatly reduces the bandwidth consumption for delivery over common paths. However, multicast MoH can be problematic in situations where a network is not enabled for multicast or where the endpoint devices are not capable of handling multicast.

There are distinct differences between unicast and multicast MoH in terms of call flow behavior. A unicast MoH call flow is initiated by a message from Unified CM to the MoH server. This message tells the MoH server to send an audio stream to the holdee device's IP address. On the other hand, a multicast MoH call flow is initiated by a message from Unified CM to the holdee device. This message instructs

the endpoint device to join the multicast group address of the configured multicast MoH audio stream. A multicast MoH server continuously streams each of the configured multicast MoH audio sources, regardless of whether any callers are on hold.

Multicast MoH is available only for IPv4. Multicast for IPv6 is not currently supported by the MoH server.

For a detailed look at MoH call flows, see the section on MoH Call Flows, page 7-23.

MoH Selection Process

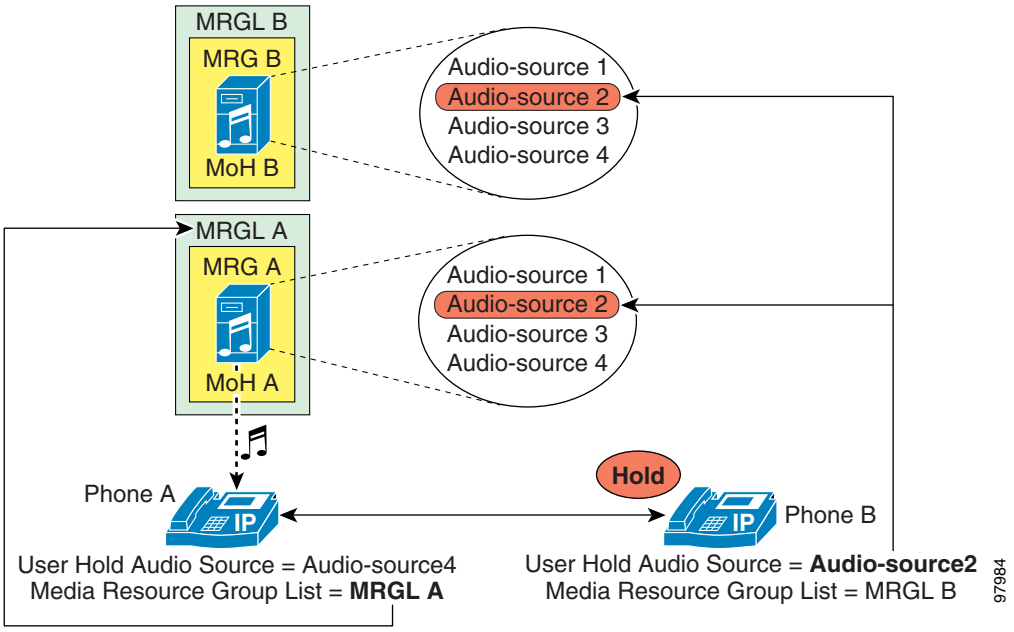
This section describes the MoH selection process as implemented in Unified CM.

The basic operation of MoH in a Cisco Unified Communications environment consists of a holder and a holdee. The *holder* is the endpoint user or network application placing a call on hold, and the *holdee* is the endpoint user or device placed on hold.

The MoH stream that an endpoint receives is determined by a combination of the User Hold MoH Audio Source of the device placing the endpoint on hold (holder) and the configured media resource group list (MRGL) of the endpoint placed on hold (holdee). The User Hold MoH Audio Source configured for the holder determines the audio file that will be streamed when the holder puts a call on hold, and the holdee's configured MRGL indicates the resource or server from which the holdee will receive the MoH stream.

As illustrated by the example in Figure 7-3, if phones A and B are on a call and phone B (holder) places phone A (holdee) on hold, phone A will hear the MoH audio source configured for phone B (Audio-source2). However, phone A will receive this MoH audio stream from the MRGL (resource or server) configured for phone A (MRGL A).

Figure 7-3 User Hold Audio Source and Media Resource Group List (MRGL)



Because the configured MRGL determines the server from which a unicast-only device will receive the MoH stream, you must configure unicast-only devices with an MRGL that points to a unicast MoH resource or media resource group (MRG). Likewise, a device capable of multicast should be configured with an MRGL that points to a multicast MRG containing a MoH server configured for multicast.

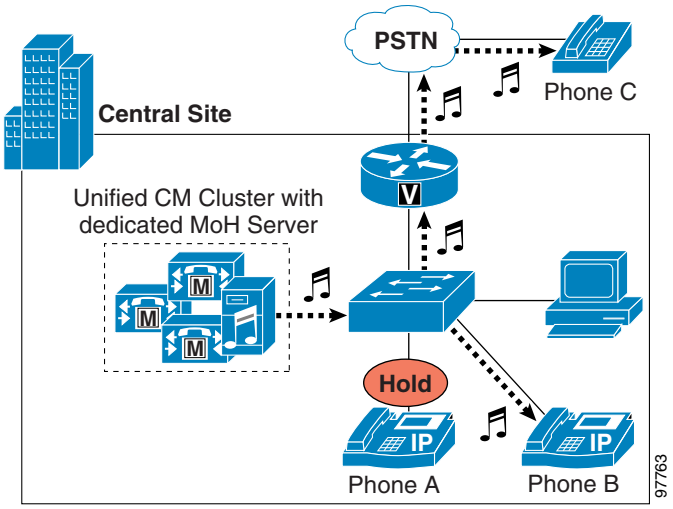
User and Network Hold

User hold includes the following types:

- User on hold at an IP phone or other endpoint device
- User on hold at the PSTN, where MoH is streamed to the gateway

Figure 7-4 shows these two types of call flows. If phone A is in a call with phone B and phone A (holder) pushes the Hold softkey, then a music stream is sent from the MoH server to phone B (holdee). The music stream can be sent to holdees within the IP network or holdees on the PSTN, as is the case if phone A places phone C on hold. In the case of phone C, the MoH stream is sent to the voice gateway interface and converted to the appropriate format for the PSTN phone. When phone A presses the Resume softkey, the holdee (phone B or C) disconnects from the music stream and reconnects to phone A.

Figure 7-4 Basic User Hold Example



Network hold can occur in following scenarios:

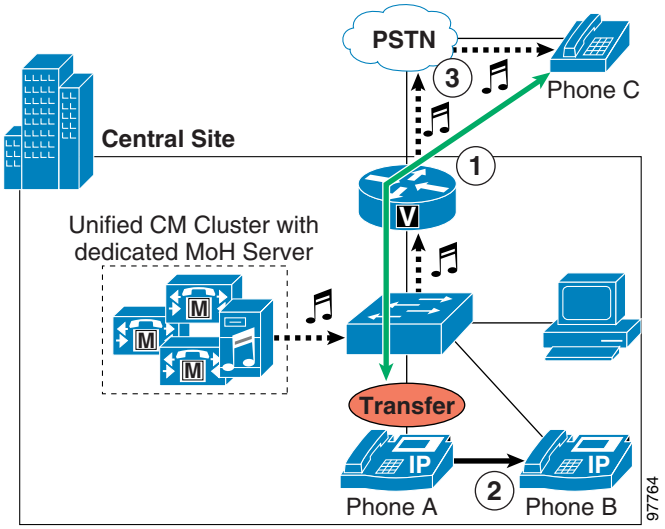
- Call transfer
- Call Park
- Conference setup
- Application-based hold

Figure 7-5 illustrates an example of network hold during a call transfer. The call flow involves the following steps:

1. Phone A receives a call from PSTN phone C.
2. Phone A answers the call and then transfers it to phone B. During the transfer process, phone C is put on network hold.
3. Phone C receives an MoH stream from the MoH server via the gateway. After phone A completes the transfer action, phone C disconnects from the music stream and gets redirected to phone B.

This process is the same for other network hold operations such as call park and conference setup.

Figure 7-5 Basic Network Hold Example for Call Transfer



MoH Sources

A Unified CM MoH server can generate a MoH stream from two types of sources:

- An audio file that has been uploaded to a Unified CM MoH server.
- A live fixed audio source or audio file from a Cisco IOS router or third-party device supporting multicast. Unified CM supports use of an external multicast audio source to provide live music on hold from external audio sources such as a CD, radio, jukebox, and so forth, as a Unified CM MoH audio source.

You can configure a maximum of 501 MoH audio sources per Unified CM cluster, of which one (the 51st) is identified as a fixed live source. An MoH server registers with the Unified CM cluster to provide IPv4 or dual-mode IPv4/IPv6 media address support.

Audio File

Audio files (.wav format) can be uploaded to Unified CM, which then automatically generates MoH audio source files for the MoH codecs. Unified CM supports G711 (a-law and mu-law), G.729 Annex A, and Cisco L16 Wideband codecs for MoH streams. The uploaded audio file(s) should be in 16-bit PCM format or 8-bit G.711 (a-law/mu-law) format.

Note Before configuring a MoH audio source, you must upload the .wav formatted audio source file to every MoH server within the cluster using the upload file function in the Unified CM Administration interface. Cisco recommends that you first upload the audio source file onto each MoH server in the cluster, then upload it onto the publisher (even if not an MoH server), and finally assign an MoH Audio Stream Number and configure the MoH audio source in the Unified CM Administration interface on the publisher. This ensures that each MoH server has the MoH audio file available when it is assigned to an MoH Audio Stream Number.

Fixed Source

If recorded or live audio is needed, multicast MoH can be generated from a fixed live source connected to the analog interface of a Cisco IOS router or third-party device that supports multicast.

Note Cisco Unified CM no longer supports a USB sound card for fixed live audio source connection to an MoH server due to lack of USB port support for MoH when Unified CM nodes are virtualized.

This mechanism enables you to use radios, CD players, or any other compatible sound source to stream multicast MoH. The stream from the fixed audio source is transcoded in real-time by the Cisco IOS router.

Note Prior to using a fixed audio source to transmit music on hold, you should consider the legalities and the ramifications of re-broadcasting copyrighted audio materials. Consult your legal department for potential issues.

For more information on live MoH from a Cisco IOS router, refer the section on *MoH from a Live Feed* in the latest version of the *Cisco Unified SCCP and SIP SRST System Administrator Guide*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps2169/products_installation_and_configuration_guides_list.html

Rebroadcast External Multicast Source

Beginning with Cisco Unified CM 11.5, Cisco IOS routers can be configured to provide a multicast audio RTP stream from a .wav file or externally connected audio source. The audio source can be from a .wav file, CD player, radio, or other audio device connected to the Cisco IOS router.

Note This feature does not affect the existing Unified CM "fixed audio source" (#51). That source remains unchanged and can be used as a fixed live source or audio file multicast streamed from a Cisco IOS router.

This feature provides the capability to configure one or more external multicast source(s) within the Unified CM MOH audio source configuration option. With this feature, Unified CM receives one or more multicast RTP streams from a Cisco IOS router with external audio source connected. In turn Unified CM unicasts the received multicast RTP streams to held callers.

Within Unified CM MoH audio sources configuration, instead of configuring an MoH audio source to use a .wav file as the audio source, you can assign an external (multicast) IP: PORT to be used as the audio source. This allows callers to hear audio on hold streamed from an external multicast source connected to the E&M port of the Cisco IOS router. (See Figure 7-6.) More than one Unified CM MoH audio source may use the same external multicast audio source.

Figure 7-6 External Multicast MoH Source

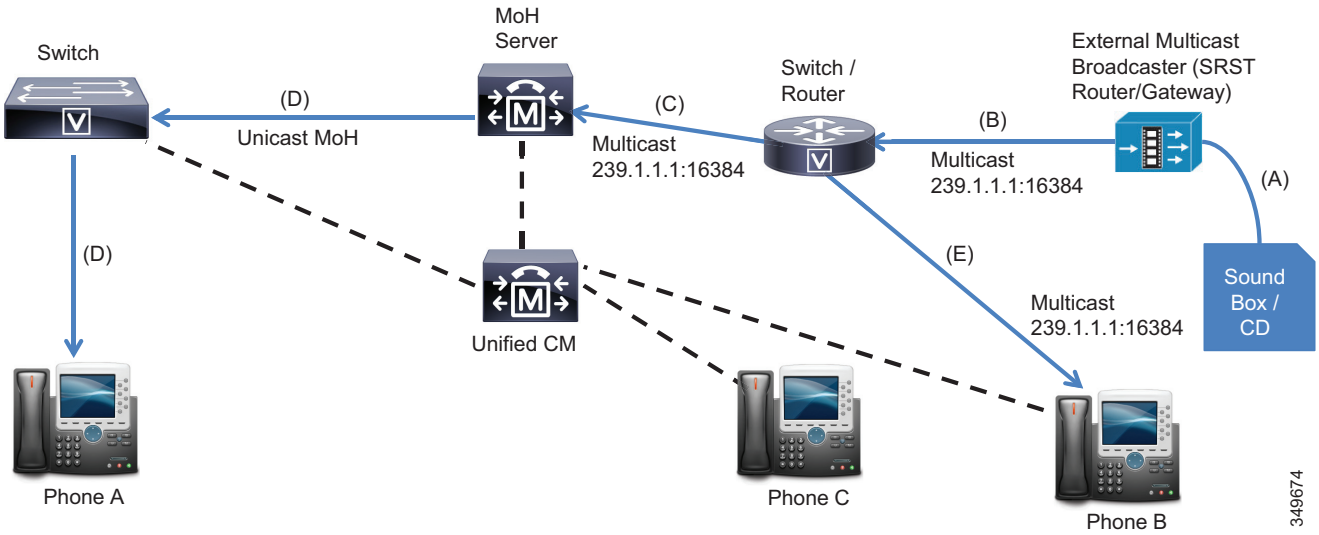


Figure 7-6 shows the network flow when using external multicast MoH sources. As shown, the music source is connected (A) to a Cisco SRST Router's E&M port, which is configured to broadcast the audio to the network multicast group 239.1.1.1:16384 (B). The MoH Server is configured to receive the multicast group (C) and rebroadcast a unicast RTP MoH stream (D) to the held Phone A.

Optionally, the MoH server could have a configured audio source with the same multicast group address that the Cisco SRST router is using to broadcast the external audio source. If the multicast MoH server and the audio source are configured with the same multicast group address, when Phone B is placed on hold, it will receive the multicast RTP audio stream (E) from the original multicast stream (B) broadcast by the SRST router. In this case the MoH server does not send the audio because it is aware that the destination multicast IP address group is the same as the external audio source multicast stream (B) broadcast by the SRST router. It is also possible to specify a different multicast IP address for the audio source on Unified CM. In this case the MoH server would rebroadcast a separate multicast stream using the original multicast broadcast audio stream (C) it receives from the network.

MoH Selection

To determine which User and Network Audio Source configuration setting to apply in a particular case, Unified CM interprets these settings for the *holder* device in the following priority order:

1. Directory or line setting (Devices with no line definition, such as gateways, do not have this level.)
2. Device setting
3. Common Device Configuration setting
4. Cluster-wide default setting

Unified CM also interprets the MRGL configuration settings of the *holdee* device in the following priority order:

1. Device setting
2. Device pool setting
3. System default MoH resources

Note that system default MoH resources are resources that are not assigned to any MRG and they are always unicast.

MoH Call Flows

The following sections provide detailed illustrations and explanations of unicast and multicast MoH call flows for both SCCP and SIP endpoints. All call flows shown below depict MoH streaming for a Unified CM MoH server. Streaming multicast MoH from Cisco IOS routers is not shown, but the call flows in those multicast scenarios are generally the same as the multicast call flows for Unified CM MoH servers.

SCCP Call Flows

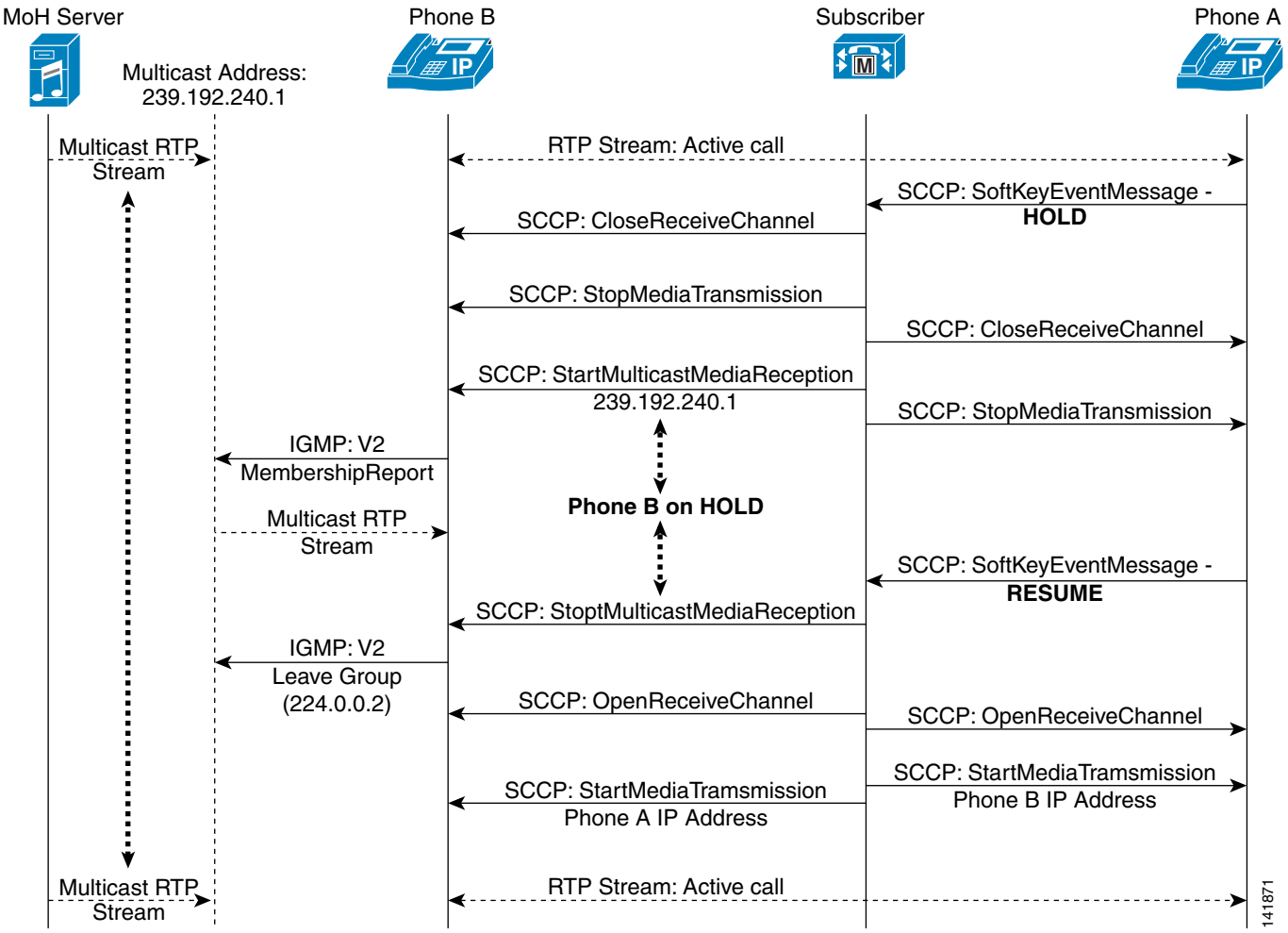
This section describes the multicast and unicast call flows for music on hold with Skinny Client Control Protocol (SCCP) endpoints.

SCCP Multicast Call Flow

Figure 7-7 illustrates a typical SCCP multicast call flow. As shown in the diagram, when the Hold softkey is pressed at phone A, Unified CM instructs both phone A and phone B to Close Receive Channel and Stop Media Transmission. This action effectively stops the RTP two-way audio stream.

Next, Unified CM tells phone B (the holdee) to Start Multicast Media Reception from multicast group address 239.192.240.1. The phone then issues an Internet Group Management Protocol (IGMP) V2 Membership Report message indicating that it is joining this group.

Figure 7-7 Detailed SCCP Multicast MoH Call Flow



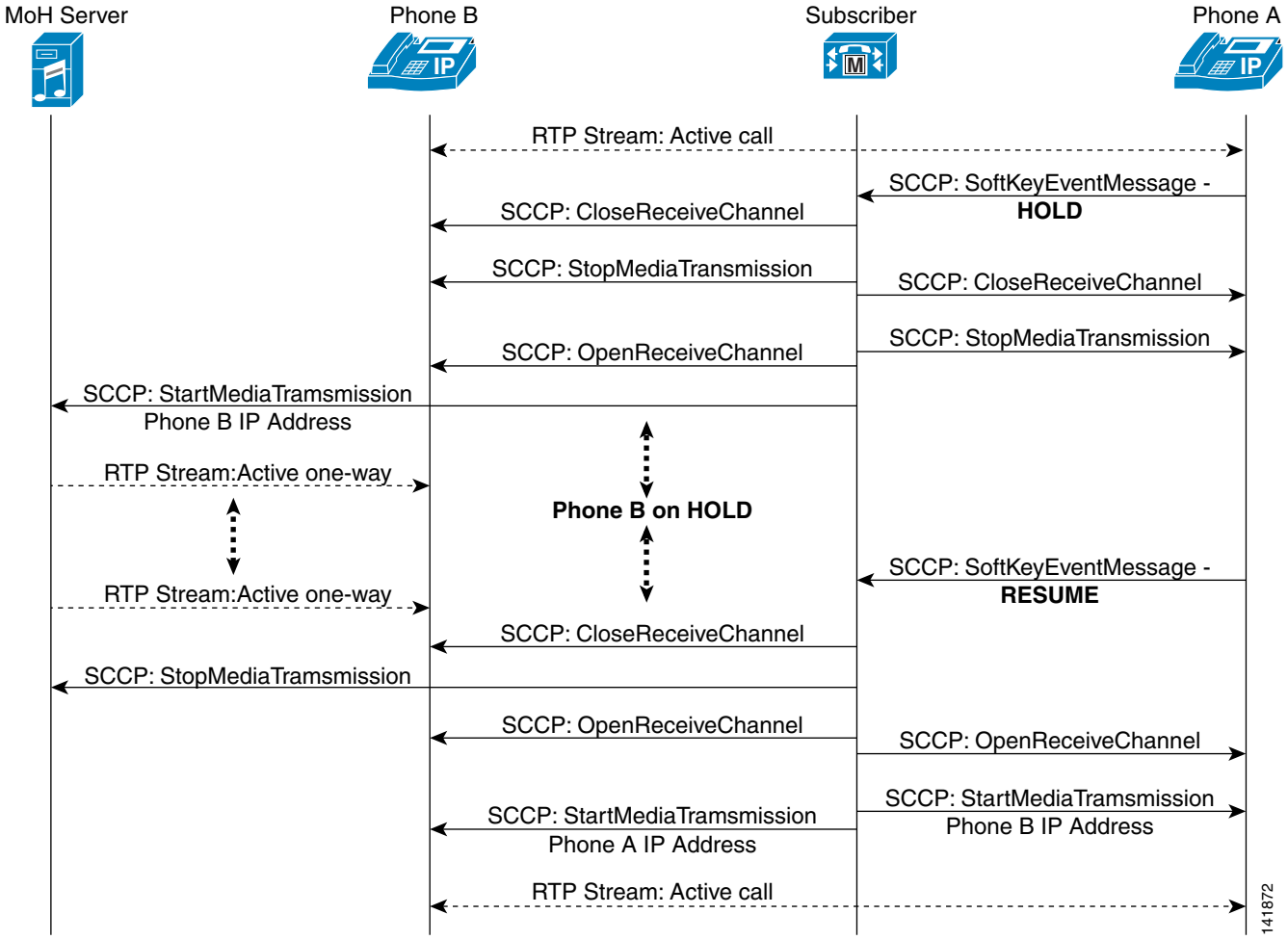
Meanwhile, the MoH server has been sourcing RTP audio to this multicast group address and, upon joining the multicast group, phone B begins receiving the MoH stream. Once phone A presses the Resume softkey, Unified CM instructs phone B to Stop Multicast Media Reception. Phone B then sends an IGMP V2 Leave Group message to 224.0.0.2 to indicate that the multicast stream is no longer needed. This effectively ends the MoH session. Next, Unified CM sends a series of Open Receive Channel messages to phones A and B, just as would be sent at the beginning of a phone call between the two phones. Soon afterwards, Unified CM instructs both phones to Start Media Transmission to each other's IP addresses. The phones are once again connected by means of an RTP two-way audio stream.

Note The call flow diagrams in Figure 7-7 and Figure 7-8 assume that an initial call exists between phones A and B, with a two-way RTP audio stream. These diagrams are representative of call flows and therefore include only the pertinent traffic required for proper MoH operation. Thus, keep-alives, acknowledgments, and other miscellaneous traffic have been eliminated to better illustrate the interaction. The initial event in each diagram is the Hold softkey action performed by phone A.

SCCP Unicast Call Flow

Figure 7-8 depicts an SCCP unicast MoH call flow. In this call flow diagram, when the Hold softkey is pressed at phone A, Unified CM instructs both phone A and phone B to Close Receive Channel and Stop Media Transmission. This action effectively stops the RTP two-way audio stream. Up to this point, unicast and multicast MoH call flows behave exactly the same way.

Figure 7-8 Detailed SCCP Unicast MoH Call Flow



Next, Unified CM tells phone B (the holdee) to Open Receive Channel. (This is quite different from the multicast case, where Unified CM tells the holdee to Start Multicast Media Reception.) Then Unified CM tells the MoH server to Start Media Transmission to the IP address of phone B. (This too is quite different behavior from the multicast MoH call flow, where the phone is prompted to join a multicast group address.) At this point, the MoH server is sending a one-way unicast RTP music stream to phone B. When phone A presses the Resume softkey, Unified CM instructs the MoH server to Stop Media Transmission and instructs phone B to Close Receive Channel, effectively ending the MoH session. As with the multicast scenario, Unified CM sends a series of Open Receive Channel messages and Start Media Transmissions messages to phones A and B with each other's IP addresses. The phones are once again connected by means of an RTP two-way audio stream.

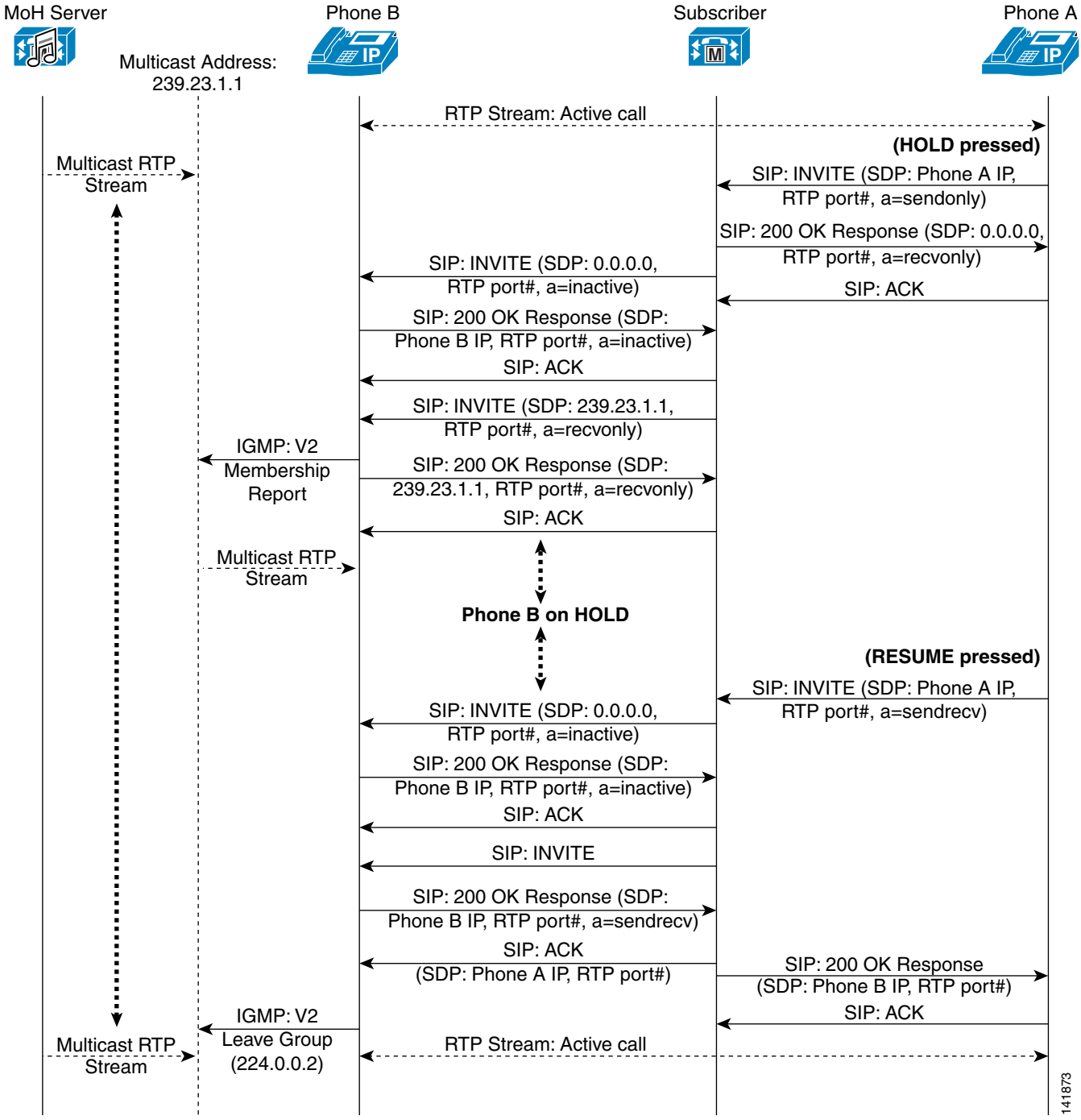
SIP Call Flows

This section describes the multicast and unicast call flows for music on hold with Session Initiation Protocol (SIP) endpoints.

SIP Multicast Call Flow

Figure 7-9 illustrates a typical SIP multicast call flow. As shown in the diagram, when the Hold softkey is pressed at phone A, phone A sends a SIP INVITE with a Session Description Protocol (SDP) connection information indication of phone A's IP address and a media attribute indication of sendonly. Unified CM then instructs phone A to disconnect the RTP stream by means of a SIP 200 OK Response with an SDP connection information indication of 0.0.0.0 and a media attribute indication of recvonly. Phone B is then told to disconnect the RTP stream by means of a SIP INVITE from Unified CM with an SDP connection information indication of 0.0.0.0 and a media attribute of inactive. After a SIP 200 OK Response is sent back from phone B to Unified CM indicating an SDP media attribute of inactive, Unified CM then sends a SIP INVITE to phone B with an SDP connection information indication of the MoH multicast group address (in this case 239.23.1.1) and a media attribute of recvonly.

Figure 7-9 Detailed SIP Multicast MoH Call Flow



Next, phone B in Figure 7-9 issues an IGMP V2 Membership Report message indicating that it is joining this multicast group. In addition, phone B sends a SIP 200 OK Response back to Unified CM indicating an SDP media attribute of recvonly in response to the previous SIP INVITE. Meanwhile, the MoH server has been sourcing RTP audio to this MoH multicast group address and, upon joining the multicast group, phone B begins receiving the one-way MoH stream.

When the user at phone A presses the Resume softkey, phone A sends a SIP INVITE with an SDP connection information indication of phone A's IP address and media attribute indications of phone A's receiving RTP port and sendrecv. Unified CM then instructs phone B to disconnect from the multicast MoH stream by means of a SIP INVITE with an SDP connection information indication of 0.0.0.0 and a media attribute indication of inactive. A SIP 200 OK Response is sent back from phone B to Unified CM, indicating an SDP media attribute of inactive.

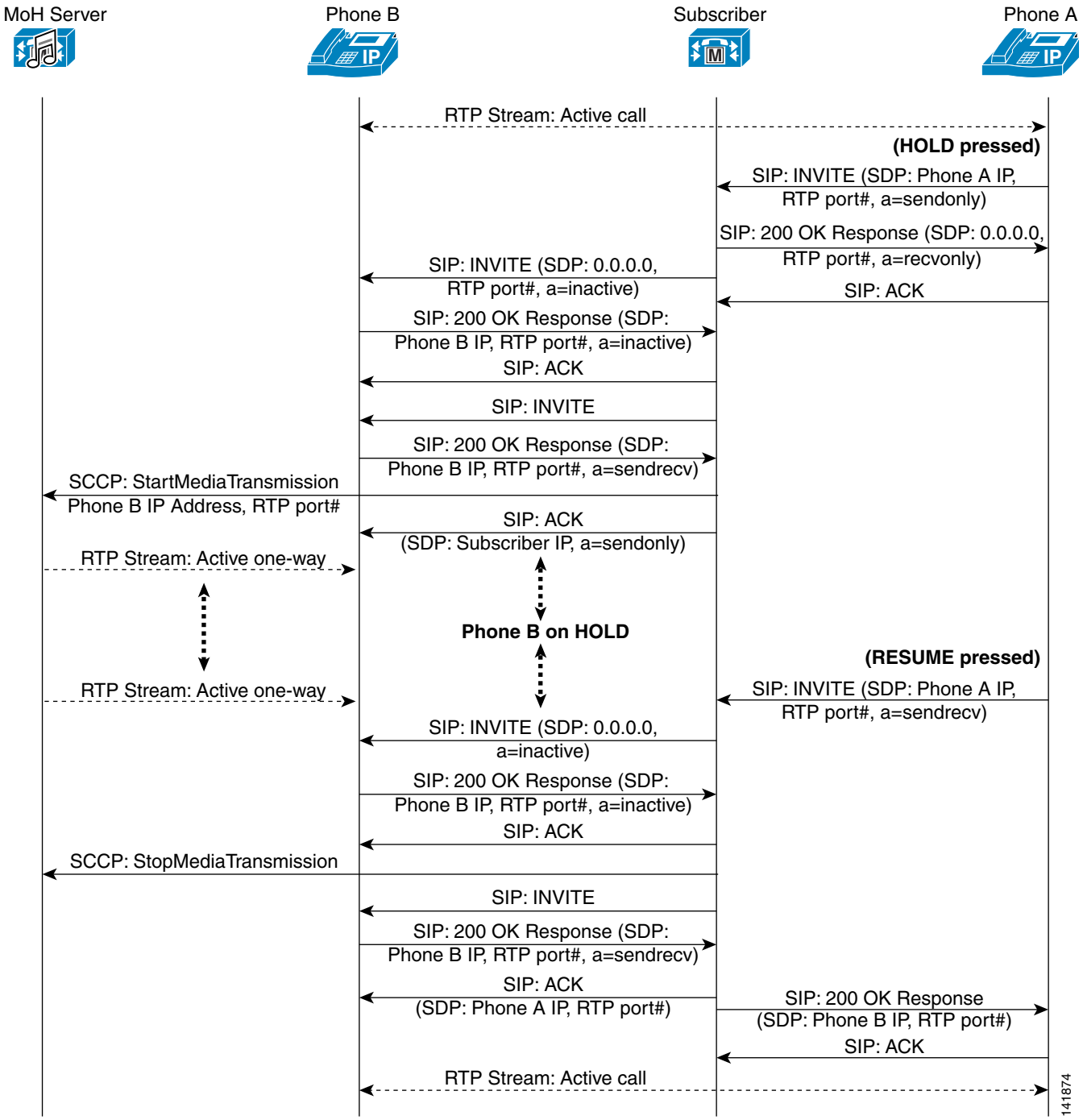
Next Unified CM sends a SIP INVITE to phone B, and phone B responds with a SIP 200 OK Response with an SDP connection information indication of phone B's IP address and media attribute indications of phone B's receiving RTP port and sendrecv. Unified CM responds by sending a SIP ACK to phone B with an SDP connection information indication of phone A's IP address and a media attribute of phone A's receiving RTP port number. Likewise, Unified CM forwards a SIP 200 OK Response to phone A's original resuming SIP INVITE, with an SDP connection information indication of phone B's IP address and a media attribute of phone B's receiving RTP port number. Phone B then sends an IGMP V2 Leave Group message to 224.0.0.2 to indicate that the multicast stream is no longer needed. Finally, the RTP two-way audio stream between phones A and B is reestablished.

Note The call flow diagrams in Figure 7-9 and Figure 7-10 assume that an initial call exists between phones A and B, with a two-way RTP audio stream. These diagrams are representative of call flows and therefore include only the pertinent traffic required for proper MoH operation. Thus, keep-alives, some acknowledgements, progression indications, and other miscellaneous traffic have been eliminated to better illustrate the interaction. The initial event in each diagram is the Hold softkey action performed by phone A.

SIP Unicast Call Flow

Figure 7-10 depicts a SIP unicast MoH call flow. As shown in the diagram, when the Hold softkey is pressed at phone A, phone A sends a SIP INVITE with an SDP connection information indication of phone A's IP address and a media attribute indication of sendonly. Unified CM then instructs phone A to disconnect the RTP stream by means of a SIP 200 OK Response with an SDP connection information indication of 0.0.0.0 and a media attribute indication of recvonly. Phone B is then told to disconnect the RTP stream by means of a SIP INVITE from Unified CM, with an SDP connection information indication of 0.0.0.0 and a media attribute of inactive. Next a SIP 200 OK Response is sent back from phone B to Unified CM, indicating an SDP media attribute of inactive. Up to this point, unicast and multicast MoH call flows are exactly the same.

Figure 7-10 Detailed SIP Unicast MoH Call Flow



Unified CM then sends a SIP INVITE to phone B, and phone B responds back with a SIP 200 OK Response indicating SDP connection information with phone B's IP address and media attribute indications of phone B's receiving RTP port number and sendrecv. Unified CM then sends a SCCP StartMediaTransmission message to the MoH server, with phone B's address and receiving RTP port

number. This is followed by a SIP ACK from Unified CM to phone B indicating SDP connection information of the Unified CM IP address and a media attribute of sendonly. Meanwhile, the MoH server begins sourcing RTP audio to phone B, and phone B begins receiving the one-way MoH stream.

When the user at phone A presses the Resume softkey, phone A sends a SIP INVITE with an SDP connection information indication of phone A's IP address and media attribute indications of phone A's receiving RTP port and sendrecv. Unified CM then instructs phone B to disconnect from the multicast MoH stream by means of a SIP INVITE with an SDP connection information indication of 0.0.0.0 and a media attribute indication of inactive. A SIP 200 OK Response is sent back from phone B to Unified CM, indicating an SDP media attribute of inactive. Then Unified CM sends an SCCP StopMediaTransmission message to the MoH server, causing the MoH server to stop forwarding the MoH stream to phone B.

Next Unified CM sends a SIP INVITE to phone B, and phone B responds with a SIP 200 OK Response with an SDP connection information indication of phone B's IP address and media attribute indications of phone B's receiving RTP port and sendrecv. Unified CM responds by sending a SIP ACK to phone B, with an SDP connection information indication of phone A's IP address and a media attribute of phone A's receiving RTP port number. Likewise, Unified CM forwards a SIP 200 OK Response to phone A's original resuming SIP INVITE with an SDP connection information indication of phone B's IP address and a media attribute of phone B's receiving RTP port. Finally, the RTP two-way audio stream between phones A and B is reestablished.

Duplex Unicast MoH Media Connections

Some scenarios require two-way media connections between held devices (holdee) and MoH servers. The Cisco Unified CM service parameter **Duplex Streaming Enabled** is available to enable this type of connection. The MoH server will discard any audio received from the held endpoints. For example, this **Duplex Streaming Enabled** option is needed when the MoH media stream must traverse through a firewall to reach the held device.

Capacity Planning for Media Resources

This section provides information on the capacities of various network modules and chassis that carry DSPs, the capacities of the chassis to carry network modules, and software dependencies of the hardware.

For all Cisco ISR G1 and G2 capacity planning, use the DSP Calculator available at https://www.cisco.com/go/dspcalculator.

The DSP resources for Unified Communications solutions are provided by NM-HD, NM-HDV, and PVDM modules. NM-HD and NM-HDV2 modules are supported on Cisco ISR G1 and G2 Series platforms. Refer to the respective product data sheets for capacity information for these modules.

PVDM modules are available in four models: PVDM-256K, PVDM2, PVDM3, and PVDM4. Each of the models has several modules with different density support.

Some things to consider when doing capacity planning for hardware-based media resources include the density of the module, the underlying platform (Cisco ISR G1 or G2), and the minimum Cisco IOS version required.

For capacity information on PVDM2 modules, refer to the *High-Density Packet Voice Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

https://www.cisco.com/en/US/prod/collateral/routers/ps5854/product_data_sheet0900aecd8016e845_ps3115_Products_Data_Sheet.html

For capacity information on PVDM3 modules, refer to the PVDM3 provisioning information available at

https://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-ip-service-level-agreements-slas/whitepaper_C11-718333.html

For capacity information on PVDM4 modules, refer to the *Cisco Fourth-Generation Packet Voice Digital Signal Processor Module for Cisco Unified Communications Solutions Data Sheet*, available at

https://www.cisco.com/c/en/us/products/routers/4000-series-integrated-services-routers-isr/datasheet-listing.html

Capacity Planning for Music on Hold

It is important to be aware of the hardware capacity for MoH resources and to consider the implications of multicast and unicast MoH in relation to this capacity when doing capacity planning for MoH resources. The capacity of the MoH server depends on several factors such as deployment model (co-resident or standalone), underlying server platform, and so forth.

Co-resident and Standalone MoH

The MoH feature requires the use of a server that is part of a Unified CM cluster. You can configure the MoH server in either of the following ways:

- Co-resident deployment

  The term *co-resident* refers to two or more services or applications running on the same server. In a co-resident deployment, the MoH feature runs on any server (either publisher or subscriber) in the cluster that is also running the Unified CM software.

- Standalone deployment

  A standalone deployment, places the MoH feature on a dedicated media resource server node within the Unified CM cluster. This server acts as neither a publisher or a subscriber. That is, the Cisco IP Voice Media Streaming Application service is the only service enabled on the server. The only function of this dedicated server is to send MoH streams to devices within the network.

Server Platform Limits

Cisco Unified Communications Manager supports a maximum of 1,000 MoH streams with Cisco Unified Computing System (UCS) C-Series or B-Series using the 7.5K or 10K Open Virtualization Archive (OVA) template for standalone deployments. For other platforms, Unified CM can support half that amount or less, depending upon what other services are active on the server. Ensure that network call volumes do not exceed these limits because, once MoH sessions have reached these limits, additional load could result in poor MoH quality, erratic MoH operation, or even loss of MoH functionality. Note that you can configure a maximum of 500 unique audio sources per Unified CM cluster.

For more information on supported MoH audio sources and sessions with each server platform, refer to the section on Media Resources, page 25-28, in the chapter on Collaboration Solution Sizing Guidance, page 25-1.

The following two MoH Server Configuration parameters affect MoH server capacity:

- **Maximum Half Duplex Streams**

  This parameter determines the number of devices that can be placed on unicast MoH. By default this value is set to 250.

  The Maximum Half Duplex Streams parameter should be set to the value derived from the following formula:

  (Server and deployment capacity) – ((Number of multicast MoH sources) * (Number of MoH codecs enabled))

  For example:

Cisco Unified Computing System (UCS) using 10K OVA standalone MoH server	Multicast MoH audio sources	MoH codecs enabled (G.711 mu-law and G.729)	Maximum half-duplex streams
1,000	- (12	* 2)	= **976**

  Therefore, in this example, the Maximum Half Duplex Streams parameter would be configured with a value of no more than 976. Each of the multicast MoH audio sources will have an automatic multicast RTP stream created for each enabled MoH codec.

- **Maximum Multicast Connections**

  This parameter determines the number of devices that can be placed on multicast MoH.

  The Maximum Multicast Connections parameter should be set to a number that ensures that all devices can be placed on multicast MoH if necessary. Although the MoH server can generate only a finite number of multicast streams, a large number of held devices can join each multicast stream. This parameter should be set to a number that is greater than or equal to the number of devices that might be placed on multicast MoH at any given time. Typically multicast traffic is accounted for based on the number of streams being generated; however, Unified CM maintains a count of the actual number of devices placed on multicast MoH or joined to each multicast MoH stream. Although this method is different than the way multicast traffic is normally tracked, it is important to configure this parameter appropriately.

Failure to configure these parameters properly could lead to under-utilization of MoH server resources or failure of the server to handle the network load. For details on how to configure the service parameters, refer to the *Cisco Unified Communications Manager Administration Guide*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Note The maximum limit of 1,000 sessions per MoH server applies to unicast, multicast, or simultaneous unicast and multicast RTP streams. The limit represents the recommended maximum number of MoH streams a platform can support, irrespective of the transport mechanism.

Resource Provisioning

When provisioning for co-resident or standalone MoH server configurations, network administrators should consider the type of transport mechanism used for the MoH audio streams. If using unicast MoH, each device on hold requires a separate MoH stream. However, if using multicast MoH and only a single audio source, then only a single MoH stream is required for each configured MoH codec type, no matter how many devices of that type are on hold.

For example, given a cluster with 30,000 phones and a 2% hold rate (only 2% of all endpoint devices are on hold at any given time), 600 MoH streams or sessions would be required. Given a unicast-only MoH environment, one co-resident (or standalone) MoH server running on a Cisco Unified Computing System (UCS) using the 10K OVA template would be required to handle this load.

By comparison, a multicast-only MoH environment with 36 unique MoH audio streams, for example, would require one co-resident MoH server. These 36 unique multicast streams could be provisioned in any one of the following ways:

- 36 unique audio sources streamed using a single codec
- 18 unique audio sources streamed using only 2 codecs
- 12 unique audio sources streamed using only 3 codecs
- 9 unique audio source streamed using all 4 codecs

In the preceding examples, the 2% hold rate is based on 30,000 phones and does not take into account gateways or other endpoint devices in the network that are also capable of being placed on hold. You should consider these other devices when calculating a hold rate because they could potentially be placed on hold just as the phones can.

The preceding calculations also do not provide for MoH server redundancy. If an MoH server fails or if more than 2% of the users go on hold at the same time, there are no other MoH resources in this scenario to handle the overflow or additional load. Your MoH resource calculations should include enough extra capacity to provide for redundancy. Additional MoH servers can be provisioned for redundancy or high availability as explained in the section on High Availability for Media Resources, page 7-34.

High Availability for Media Resources

The Unified CM constructs of media resource groups (MRGs) and media resource group lists (MRGLs) are used to control how the resources described in this chapter are organized and accessed. This section discusses considerations for how to utilize these constructs effectively.

Media Resource Groups and Lists

Media resource groups (MRGs) and media resource lists (MRGLs) provide a method to control how resources are allocated that could include rights to resources, location of resources, or resource type for specific applications. This section assumes you have an understanding of media resource groups and lists, and it highlights the following design considerations:

- The system defines a default media resource group that is not visible in the user interface. All resources are members of this default MRG when they are created. When using MRGs to control access to resources, it is necessary to move the resources out of the default MRG by explicitly configuring them in some other MRG. If the desired effect is for resources to be available only as a last resort for all calls, then the resources may remain in the default group. Also, if no control over resources is necessary, they may remain in the default group.
- Consumers of media resources use resources first from any media resource group (MRG) or media resource group list (MRGL) that their configuration specifies. If the required resource is not available, the default MRG is searched for the resource. For simple deployments, the default MRG alone may be used.
- Use media resource groups (MRGs) and media resource group lists (MRGLs) to provide sharing of resources across multiple Unified CMs. If you do not use MRGs and MRGLs, the resources are available to a single Unified CM only.
- MRGLs will use MRGs in the order that they are listed in the configuration. If one MRG does not have the needed resource, the next MRG is searched. If all MRGs are searched and no resource is found, the search terminates.
- Within an MRG, resources are allocated based on their order in their configuration even though Unified CM Administration displays the devices in an MRG in alphabetical order. If you want media resources to be allocated in a specific order, Cisco recommends that you create a separate MRG for each individual resource and use MRGLs to specify the order of allocation.
- When there are multiple devices providing the same type of resource within an MRG, the algorithm for allocating that resource load-balances across all those devices. Cisco Unified CM uses a throttling mechanism to load balance across MTP and transcoder resources using the **MTP and Transcoder Resource Throttling Percentage** service parameter, which defines a percentage of the configured number of MTP or transcoder resources. When the number of active MTP or transcoder resources is equal to or greater than the percentage that is configured for this parameter, Cisco Unified CM stops sending calls to this resource and hunts through the MRGL (including the default MRG) one time to find a resource that uses matching codecs on both sides of the call. If Cisco Unified CM cannot find an available resource with matching codecs, it returns to the top of the MRGL to repeat the search, which then includes those resources that are in a throttled state and that match a smaller subset of capabilities for the call. Cisco Unified CM extends the call to the resource that is the best match for the call when such a resource is available. The call fails when Cisco Unified CM cannot allocate a resource for the call.

- Unified CM server-based software MTPs are pass-through enabled by default. Cisco IOS Enhanced MTP devices can be configured to support codec pass-through or non-codec pass-through modes. If a codec pass-through MTP is required, then Unified CM looks for an MRGL (including a default MRG) to find an MTP with Real-Time Transport Control Protocol pass-through (RTCP PT) capabilities in the first iteration. If it cannot find an MTP with the requested RTCP pass-through capabilities, Unified CM will go through the MTP list again requesting for pass-through capability. If it cannot find an MTP with the requested pass-through capabilities, Unified CM will go through the MTP list again without requesting for pass-through capability.
- An MRG may contain multiple types of resources, and the appropriate resource will be allocated from the group based on the feature needed. MTPs and transcoders are a special case because a transcoder may also be used as an MTP. For example, when both MTPs and transcoders exist in the same MRG and an MTP is required, the allocation is done based on the order in which the resources appear in the MRG. If transcoder devices appear earlier than MTPs in the MRG, transcoder resources will be allocated for the MTP requirement until the transcoder resources are exhausted and then the system will start allocating MTPs. For this reason, it is important to consider the order of resources when creating MRGs and MRGLs.
- MRGs can also be used to group resources of similar types. As explained in the example above, because a transcoder is a more expensive resource, Cisco recommends grouping transcoders and MTPs into separate MRGs and invoking the right resource by adding MRGs to the MRGL in appropriate order.
- You can also use MRGs and MRGLs to separate resources based on geographical location, thereby conserving WAN bandwidth whenever possible.
- Ensure that the media resources themselves have configurations that prevent further invocation of other media resources. For example, if an MTP is inserted into a call and the codec configured on that MTP does not match the one needed by Unified CM for the call, then a transcoder may also be invoked. A frequent mistake is to configure an MTP for G.729 or G.729b when Unified CM needs G.729a.

Redundancy and Failover Considerations for Cisco IOS-Based Media Resources

A high availability design with media resources must include redundant media resources. When these resources are Cisco IOS-based, they can be distributed on more than one Cisco IOS platform to guard against failure of a single platform and they can be registered to different primary Unified CM servers.

Cisco IOS supports two modes of failover capability: graceful and immediate. The default failover method is graceful, in which the resources register to a backup Unified CM server only after all media activity has ceased. The immediate method, on the other hand, makes the resources register to the backup Unified CM server as soon as failure of the primary is detected. In situations where there is only one set of media resources with no redundancy, Cisco recommends use of the immediate failover method.

High Availability for Transcoders

The following transcoder failover process takes place in the event that the Cisco Unified CM to which the device is registered becomes unavailable:

If the primary Unified CM fails, the transcoder device attempts to register with the secondary Unified CM node as defined in the Cisco Unified CM group for that device. The transcoder device will fall back to the primary Unified CM as soon it becomes available again. The calls that were on that Unified CM will register with the next Unified CM in the list.

High Availability for Music on Hold

Cisco recommends that you configure and deploy multiple MoH servers for completely redundant MoH operation. If the first MoH server fails or becomes unavailable because it no longer has the resources required to service requests, the second server can provide continued MoH functionality. For proper redundant configuration, assign resources from at least two MoH servers to each MRG in the cluster.

In environments where both multicast and unicast MoH are required, be sure to provide redundancy for both transport types to ensure MoH redundancy for all endpoints in the network.

Design Considerations for Media Resources

This section discusses specific considerations for deploying media resources for use with the various Unified CM deployment models. It also highlights the configuration considerations and best practices to help you design a robust solution for media resource allocation in your Unified CM implementation.

Deployment Models

This section examines where and when the MTP and transcoding resources are used within the following three enterprise IP Telephony deployment models:

- Single-Site Deployments, page 7-36
- Multisite Deployments with Centralized Call Processing, page 7-37
- Multisite Deployments with Distributed Call Processing, page 7-38

Single-Site Deployments

In a single-site deployment, there is no need for transcoding because there are no low-speed links to justify the use of a low bit-rate (LBR) codec. Some MTP resources might be required in the presence of a significant number of devices that are not compliant with H.323v2, such as older versions of Microsoft NetMeeting or certain video devices. MTP resources may be required for DTMF conversion if SIP endpoints are present (see Named Telephony Events (RFC 2833), page 7-7.)

In a single-site deployment, if Unified CM receives an inbound call from an SCCP-based Cisco Unified IP Phone 7940 or 7960, the media capabilities of the calling device are not available when the call is initiated, and most of the SIP PSTN service providers require an early offer. In this case, Unified CM must insert an MTP and use its IP address and UDP port number to advertise all supported audio codecs (after region-pair filtering) in the Offer SDP of the initial INVITE sent over the outbound SIP trunk.

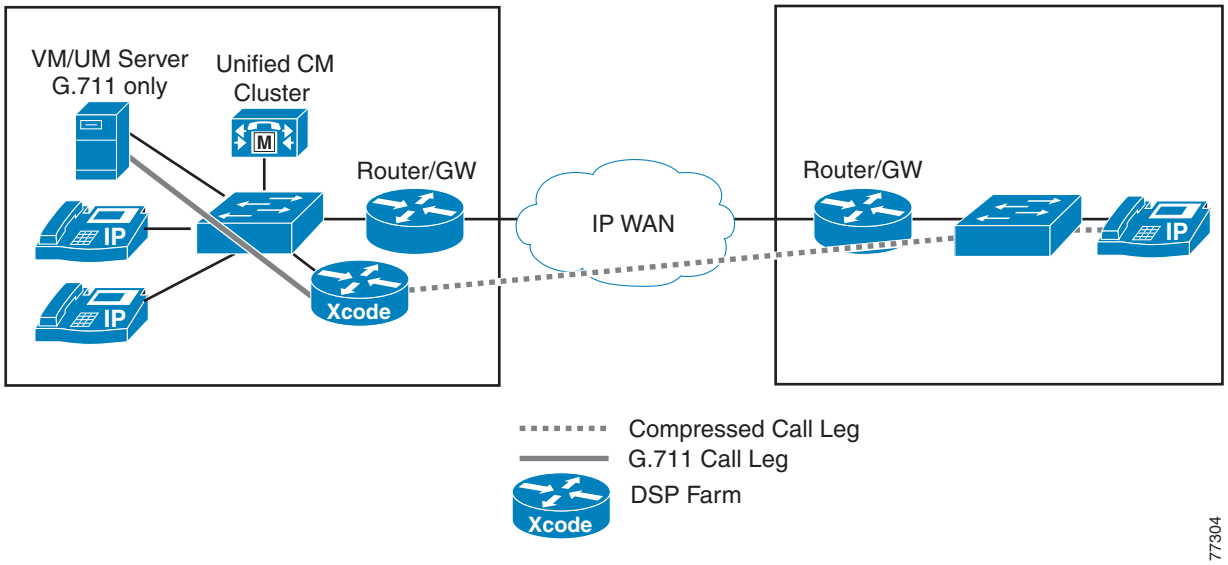
Multisite Deployments with Centralized Call Processing

In a centralized call processing deployment, the Unified CM cluster and the applications (such as voice mail and IVR) are located at the central site, while several remote sites are connected through an IP WAN. The remote sites rely on the centralized Unified CMs to handle their call processing.

Because WAN bandwidth is typically limited, calls are configured to use a low bit-rate codec such as G.729 when traversing the WAN. (See Figure 7-11.)

Voice compression between IP phones is easily configured through the use of *regions* and *locations* in Unified CM. A region defines the type of compression (for example, G.711 or G.729) used by the devices in that region, and a location specifies the total amount of bandwidth available for calls to and from devices at that location.

Figure 7-11 Transcoding for the WAN with Centralized Call Processing



Unified CM uses media resource groups (MRGs) to enable sharing of MTP and transcoding resources among the Unified CM servers within a cluster. In addition, when using an LBR codec (for example, G.729a) for calls that traverse different regions, the transcoding resources are used only if one (or both) of the endpoints is unable to use the LBR codec.

In Figure 7-11, Unified CM knows that a transcoder is required and allocates one based on the MRGL and/or MRG of the device that is using the higher-bandwidth codec. In this case it is the VM/UM server that determines which transcoder device is used. This behavior of Unified CM is based on the assumption that the transcoder resources are actually located close to the higher-bandwidth device. If this system was designed so that the transcoder for the VM/UM server was located at the remote site, then G.711 would be sent across the WAN, which would defeat the intended design. As a result, if there are multiple sites with G.711-only devices, then each of these sites would need transcoder resources when an LBR is run on the WAN.

The placement of other resources is also important. For example, if a conference occurs with three phones at a remote site and the conference resource is located in the central (call processing) site, then three media streams are carried over the WAN. If the conference resource were local, then the calls would not traverse the WAN. It is necessary to consider this factor when designing the bandwidth and call admission control for your WAN.

Multisite Deployments with Distributed Call Processing

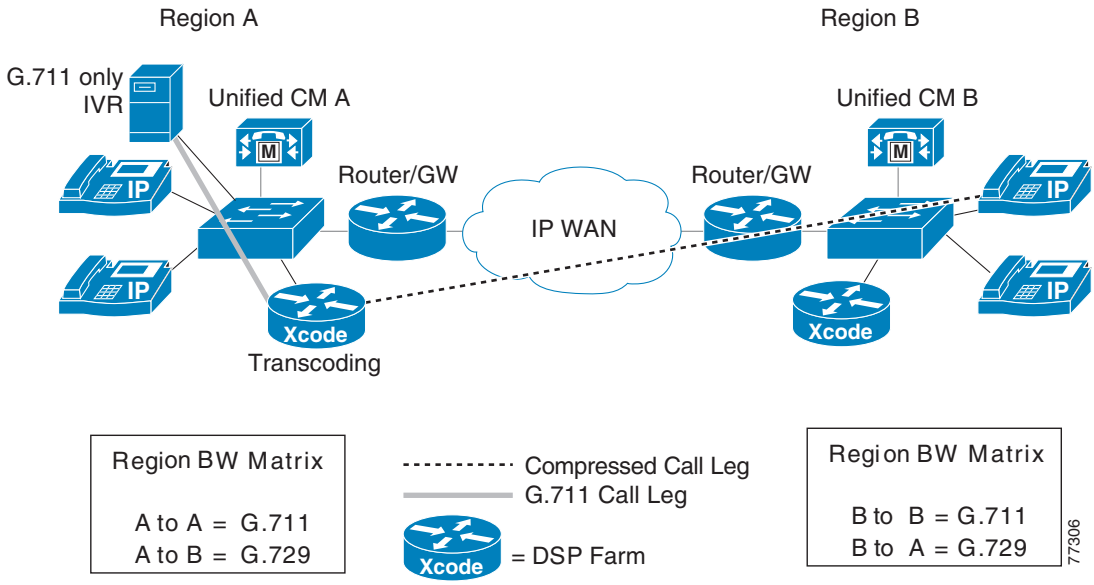
In distributed call processing deployments, several sites are connected through an IP WAN. Each site contains a Unified CM cluster that can, in turn, follow the single-site model or the centralized call processing model. A gatekeeper may be used for call admission control between sites.

Because WAN bandwidth is typically limited, calls between sites may be configured to use an LBR codec (such as G.729a) when traversing the WAN. H.323v2 intercluster trunks are used to connect Unified CM clusters. Unified CM also supports compressed voice call connections through the MTP service if a hardware MTP is used. (See Figure 7-12.)

A distributed call processing deployment might need transcoding and MTP services in the following situations:

- With current versions of Cisco applications, it is possible and recommended to avoid the use of transcoding resources. There might be specific instances where G.711 on a specific device cannot be avoided.
- Some endpoints (for example, video endpoints) do not support the H.323v2 features.

Figure 7-12 Intercluster Call Flow with Transcoding



Unified CM uses media resource groups (MRGs) to enable sharing of MTP and transcoding resources among the Unified CM servers within a cluster. In addition, for calls across intercluster trunks, MTP and transcoding resources are used only when needed, thus eliminating the need to configure the MTP service for applications that do not support LBR codecs.

The following characteristics apply to distributed call processing deployments:

- Only the intercluster calls that require transcoding will use the MTP service. For example, if both endpoints of a call are capable of using a G.729 codec, no transcoding resources will be used.
- Sharing MTP resources among servers within a cluster provides more efficient resource utilization.

Media Functions and Voice Quality

Any process that manipulates media can degrade the quality of the media. For example, encoding a voice stream for transmission across any network (IP or TDM) and decoding it at the other end will result in a loss of information, and the resulting voice stream will not be an exact reproduction of the original. If there are media traversal paths through the network that involve multiple encoding and decoding steps of the same voice stream, then each successive encoding/decoding operation will further degrade the voice quality. In general, such paths should be avoided. This is especially true for low-bandwidth codecs (LBC) such as G.729.

If such paths cannot be avoided, voice quality can generally be improved by using a higher bandwidth, low-compression codec, such as the G.711 or G.722 codecs, which are recommended wherever such paths are anticipated. Use of lower bandwidth, higher compression codecs in such scenarios is not recommended.

Music on Hold Design Considerations

This section highlights some MoH configuration considerations and best practice to help you design a robust MoH solution.

Codec Selection

If you need multiple codecs for MoH deployment, configure them in the IP Voice Media Streaming Application service parameter **Supported MoH Codecs** under the Clusterwide Unified CM Service Parameters Configuration. From the Supported MoH Codecs list under the Clusterwide Parameters, select all the desired codec types that should be allowed for MoH streams. By default, only G.711 mu-law is selected. To select another codec type, click on it in the scrollable list. For multiple selections, hold down the CTRL key and use the mouse to select multiple codecs from the scrollable list. The actual codec used for a MoH event is determined by the Region settings of the MoH server and the device being put on hold (IP phone, gateway, and so forth). Therefore, assign the proper Region setting to your MoH servers and configure the desired Region Relationships to control the codec selection of MoH interactions.

Note If you are using the G.729 codec for MoH audio streams, be aware that this codec is optimized for speech and it provides only marginal audio fidelity for music.

Multicast Addressing

Proper IP addressing is important for configuring multicast MoH. Addresses for IP multicast range from 224.0.1.0 to 239.255.255.255. The Internet Assigned Numbers Authority (IANA), however, assigns addresses in the range 224.0.1.0 to 238.255.255.255 for public multicast applications. Cisco strongly discourages using public multicast addresses for music on hold. Instead, Cisco recommends that you configure multicast MoH audio sources to use IP addresses in the range 239.1.1.1 to 239.255.255.255, which is reserved for administratively controlled applications on private networks.

Furthermore, you should configure multicast audio sources to increment on the IP address and not the port number, for the following reasons:

- IP phones placed on hold join multicast IP addresses, not port numbers.

  Cisco IP phones have no concept of multicast port numbers. Therefore, if all the configured codecs for a particular audio stream transmit to the same multicast IP address (even on different port numbers), all streams will be sent to the IP phone even though only one stream is needed. This has the potential of saturating the network with unnecessary traffic because the IP phone is capable of receiving only a single MoH stream.

- IP network routers route multicast based on IP addresses, not port numbers.

  Routers have no concept of multicast port numbers. Thus, when it encounters multiple streams sent to the same multicast group address (even on different port numbers), the router forwards all streams of the multicast group. Because only one stream is needed, network bandwidth is over-utilized and network congestion can eventually result.

When configuring multiple multicast MoH servers, assign a different base multicast IP address and/or range to each MoH server. If multiple MoH servers are transmitting to the same multicast IP address, then when an endpoint joins the multicast group address, it will receive multiple MoH streams (from different MoH servers).

Unified CM MoH Audio Sources

Configured audio sources are shared among *all* MoH servers in the Unified CM cluster, requiring each audio source file to be uploaded to every MoH server within the cluster. You can configure up to 500 unique audio sources per cluster.

For those audio sources that will be used for multicast streaming, ensure that **Allow Multicasting** is enabled.

Unicast and Multicast in the Same Unified CM Cluster

In some cases, administrators might want to configure a single Unified CM cluster to handle both unicast and multicast MoH streams. This configuration might be necessary because the telephony network contains devices or endpoint that do not support multicast or because some portions of the network are not enabled for multicast.

Use one of the following methods to enable a cluster to support both unicast and multicast MoH audio streams:

- Deploy separate MoH servers, with one server configured as a unicast MoH server and the second server configured as a multicast MoH server.
- Deploy a single MoH server with two media resource groups (MRGs), each containing the same MoH server, with one MRG configured to use multicast for audio streams and the second MRG configured to use unicast.

In either case, you must configure at least two MRGs and at least two media resource group lists (MRGLs). Configure one unicast MRG and one unicast MRGL for those endpoints requiring unicast MoH. Likewise, configure one multicast MRG and one multicast MRGL for those endpoints requiring multicast MoH.

When deploying separate MoH servers, configure one server without multicast enabled (unicast-only) and configure a second MoH server with multicast enabled. Assign the unicast-only MoH media resource and the multicast-enabled MoH media resource to the unicast and multicast MRGs, respectively. Ensure that the **Use Multicast for MoH Audio** box is checked for the multicast MRG but

not for the unicast MRG. Also assign these unicast and multicast MRGs to their respective MRGLs. In this case, an MoH stream is unicast or multicast based on whether the MRG is configured to use multicast and then on the server from which it is served.

When deploying a single MoH server for both unicast and multicast MoH, configure the server for multicast. Assign this same MoH media resource to both the unicast MRG and the multicast MRG, and check the **Use Multicast for MoH Audio** box for the multicast MRG. In this case, an MoH stream is unicast or multicast based solely on whether the MRG is configured to use multicast.

Note When configuring the unicast MRG, do not be confused by the fact that the MoH media resource you are adding to this MRG has [Multicast] appended to the end of the resource name even though you are adding it to the unicast MRG. This label is simply an indication that the resource is capable of being multicast, but the **Use Multicast for MoH Audio** box determines whether the resource will use unicast or multicast.

In addition, you must configure individual devices or device pools to use the appropriate MRGL. You can place all unicast devices in a device pool or pools and configure those device pools to use the unicast MRGL. Likewise, you can place all multicast devices in a device pool or pools and configure those device pools to use the multicast MRGL. Optionally, you can configure individual devices to use the appropriate unicast or multicast MRGL. Lastly, configure a User Hold Audio Source and Network Hold Audio Source for each individual device or (in the case of phone devices) individual lines or directory numbers to assign the appropriate audio source to stream.

When choosing a method for deploying both multicast and unicast MoH in the same cluster, an important factor to consider is the number of servers required. When using a single MoH server for both unicast and multicast, fewer MoH servers are required throughout the cluster. Deploying separate multicast and unicast MoH servers will obviously require more servers within the cluster.

Quality of Service (QoS)

Convergence of data and voice on a single network requires adequate QoS to ensure that time-sensitive and critical real-time applications such as voice are not delayed or dropped. To ensure proper QoS for voice traffic, the streams must be marked, classified, and queued as they enter and traverse the network to give the voice streams preferential treatment over less critical traffic. MoH servers automatically mark audio stream traffic the same as voice bearer traffic, with a Differentiated Services Code Point (DSCP) value of 46 or a Per Hop Behavior (PHB) value of EF (ToS of 0xB8). Therefore, as long as QoS is properly configured on the network, MoH streams will receive the same classification and priority queueing treatment as voice RTP media traffic.

Call signaling traffic between MoH servers and Unified CM servers is automatically marked with a DSCP value of 24 or a PHB value of CS3 (ToS of 0x60) by default. Therefore, as long as QoS is properly configured on the network, this call signalling traffic will be properly classified and queued within the network along with all other call signalling traffic.

Call Admission Control and MoH

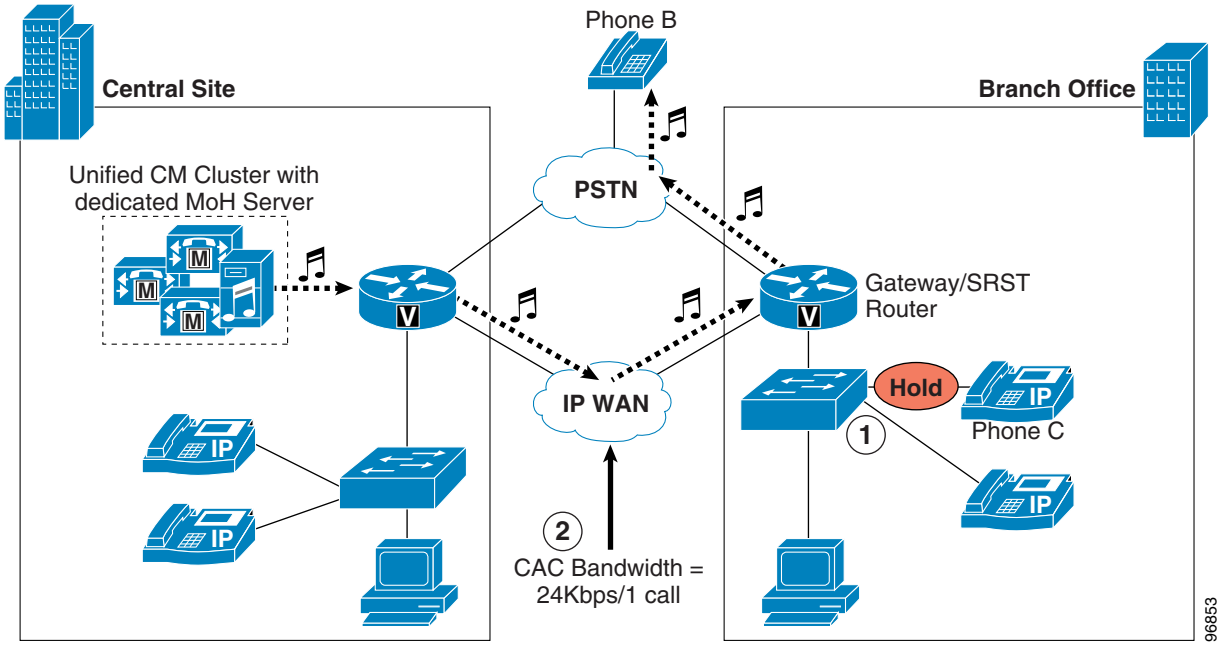
Call admission control (CAC) is required when IP telephony traffic is traveling across WAN links. Due to the limited bandwidth available on these links, it is highly probable that voice media traffic might get delayed or dropped without appropriate call admission control. For additional information, see Bandwidth Management, page 13-1.

Call admission control for Unified CM (based on either static locations or RSVP-enabled locations) is capable of tracking unicast MoH streams traversing the WAN but not multicast MoH streams. Thus, even if WAN bandwidth has been fully subscribed, a multicast MoH stream will not be denied access to the WAN by call admission control. Instead, the stream will be sent across the WAN, likely resulting in poor audio stream quality and poor quality on all other calls traversing the WAN. To ensure that multicast MoH streams do not cause this over-subscription situation, you should over-provision the QoS configuration on all downstream WAN interfaces by configuring the low-latency queuing (LLQ) voice priority queue with additional bandwidth. Because MoH streams are uni-directional, only the voice priority queues of the downstream interfaces (from the central site to remote sites) must be over-provisioned. Add enough bandwidth for every unique multicast MoH stream that might traverse the WAN link. For example, if there are four unique multicast audio streams that could potentially traverse the WAN, then add 96 kbps to the voice priority queue (4 * 24 kbps per G.729 audio stream = 96 kbps).

Figure 7-13 shows an example of call admission control and MoH in a centralized multisite deployment. For this example, assume that IP phone C is in a call with a PSTN phone (phone B). At this point, no bandwidth has been consumed on the WAN. When phone C pushes the Hold softkey (step 1), phone B receives an MoH stream from the central-site MoH server by way of the WAN, thereby consuming bandwidth on the link. Whether or not this bandwidth is taken into consideration by call admission control depends on the type of MoH stream. If multicast MoH is streamed, then call admission control will not consider the 24 kbps being consumed (therefore, QoS on the downstream WAN interfaces should be provisioned accordingly). However, if unicast MoH is streamed, call admission control will subtract 24 kbps from the available WAN bandwidth (step 2).

Note The preceding example might seem to imply that unicast MoH should be streamed across the WAN. However, this is merely an example used to illustrate locations-based call admission control with MoH and is not intended as a recommendation or endorsement of this configuration. As stated earlier, multicast MoH is the recommended transport mechanism for sending MoH audio streams across the WAN.

Figure 7-13 Locations-Based Call Admission Control and MoH



Deployment Models for Music on Hold

The various Unified Communications call processing deployment models introduce additional considerations for MoH configuration design. Which deployment model you choose can also affect your decisions about MoH transport mechanisms (unicast or multicast), resource provisioning, and codecs. This section discusses these issues in relation to the various deployment models.

For more detailed information about the deployment models, see the chapter on Collaboration Deployment Models, page 10-1.

Single-Site Campus (Relevant to All Deployments)

Single-site campus deployments are typically based on a LAN infrastructure and provide sufficient bandwidth for large amounts of traffic. Because bandwidth is typically not limited in a LAN infrastructure, Cisco recommends the use of the G.711 (A-law or mu-law) codec for all MoH audio streams in a single-site deployment. G.711 provides the optimal voice and music streaming quality in an IP Telephony environment.

MoH server redundancy should also be considered. In the event that an MoH server becomes overloaded or is unavailable, configuring multiple MoH servers and assigning them in preferred order to MRGs ensures that another server can take over and provide the MoH streams.

With the increasing diversity of network technologies, in a large single-site campus it is likely that some endpoint devices or areas of the network will be unable to support multicast. For this reason, you might have to deploy both unicast and multicast MoH resources. For more information, see the section on Unicast and Multicast in the Same Unified CM Cluster, page 7-40.

To ensure that off-net calls and application-handled calls receive expected MoH streams when placed on hold, configure all gateways and other devices with the appropriate MRGLs and audio sources, or assign them to appropriate device pools.

Centralized Multisite Deployments

Multisite IP telephony deployments with centralized call processing typically contain WAN connections to multiple non-central sites. These WAN links usually cause bandwidth and throughput bottlenecks. To minimize bandwidth consumption on these links, Cisco recommends the use of the G.729 codec for all MoH audio streams traversing the WAN. Because the G.729 codec is optimized for voice and not music applications, you should use G.729 only across the WAN, where the bandwidth savings far outweighs the lower quality afforded by G.729 for MoH transport. Likewise, because multicast traffic provides significant bandwidth savings, you should always use multicast MoH when streaming audio to endpoints across the WAN.

If the sound quality of an MoH stream becomes an issue when using the G.729 codec across the WAN, you can use the G.711 codec for MoH audio streams across the WAN while still using G.729 for voice calls. In order to send MoH streams across the WAN with the G.711 codec but voice calls across the WAN with the G.729 codec, place all MoH servers in a Unified CM region by themselves, and configure that region to use G.711 between itself and all other regions. Thus, when a call is placed between two phones on either side of a WAN, the G.729 codec is used between their respective regions. However, when the call is placed on hold by either party, the MoH audio stream is encoded using G.711 because G.711 is the configured codec to use between the MoH server's region and the region of the phone placed on hold.

Centralized PSTN Deployments

In a centralized PSTN deployment with a single gateway or set of gateways for PSTN access, there is no way to configure more than 500 unique audio sources. The media resource group list (MRGL) assigned to the gateway determines the MoH server used for MoH streaming when a PSTN caller is placed on hold, while the phone placing the call on hold determines the audio source. With a centralized PSTN deployment, because there are no PSTN gateways at the local sites, an MRGL cannot be used to point to multiple MoH servers based on branch location. Thus, in this case there are at most 500 unique site-specific MoH sources or multicast streaming addresses to enable unique MoH sources for a maximum of 500 branch sites.

The following example illustrates how MoH streaming works for up to 500 locations:

For branches 1 to 500, the centralized PSTN gateway or set of gateways is/are configured with MRGL pointing to MoH server node MoH_1 (with base multicast address of 239.1.1.1), with all phones at each branch pointing to one of the 500 audio sources configured on the cluster. Thus, phones in Branch 1 point to audio source 1, which on MoH_1 server is 239.1.1.1 to 4 (depending on codec and assuming audio sources are configured in order); phones in Branch 2 point to audio source 2, which on MoH_1 server is 239.1.1.5 to 8; phones in Branch 3 point to audio source 3, which on MoH_1 server is 239.1.1.9 to 12; and so on up to phones in Branch 500 pointing to audio source 501, which on MoH_1 server is 239.1.8.197 to 200.

Multicast MoH from Branch Routers

Branch routers deployed with the Cisco Unified Survivable Remote Site Telephony (SRST) feature can provide multicast MoH in a remote or branch site, with the MoH streaming from the branch SRST router's flash or from a live feed connected to an analog port. Multicast MoH from a branch router via these two methods enhances the Unified CM MoH feature in both of the following scenarios:

- Non-Fallback Mode

  When the WAN is up and the phones are controlled by Unified CM, this configuration can eliminate the need to forward MoH across the WAN to remote branch sites by providing locally sourced MoH.

- Fallback Mode

  When SRST is active and the branch devices have lost connectivity to the central-site Unified CM, the branch router can continue to provide multicast MoH.

When using the live feed option in either scenario, the SRST router provides redundancy by monitoring the live feed input, and it will revert to streaming MoH from a file in flash if the live feed is disconnected. You can use only a single multicast address and port number per SRST router to provide multicast MoH; therefore, the SRST router does not support streaming from both the live feed and the flash file at the same time. In addition, the SRST router can stream only a single audio file from flash.

Note An SRST license is required regardless of whether the SRST functionality will actually be used. The license is required because the configuration for streaming MoH from branch router flash is done under the SRST configuration mode and, even if SRST functionality will not be used, at least one **max-ephones** and one **max-dn** must be configured.

Non-Fallback Mode

During non-fallback mode (when the WAN is up and SRST is not active), the branch SRST or E-SRST router can provide multicast MoH to all local Cisco Unified Communications devices. To accomplish this, you must configure a Unified CM MoH server with an audio source that has the same multicast IP

address and port number as configured on the branch router. The audio source multicast IP address and port number used on the branch router can correspond to the multicast address and port number of either an audio source file or the fixed audio source of the centralized Unified CM MoH server. In this scenario, because the multicast MoH audio stream is always coming from the SRST or E-SRST router, it is not necessary for the central-site MoH server audio source to traverse the WAN.

To prevent the central-site audio stream(s) from traversing the WAN, use one of the following methods:

- Configure a maximum hop count

  Configure the central-site MoH audio source with a maximum hop count (or TTL) low enough to ensure that it will not stream further than the central-site LAN.

- Configure an access control list (ACL) on the WAN interface

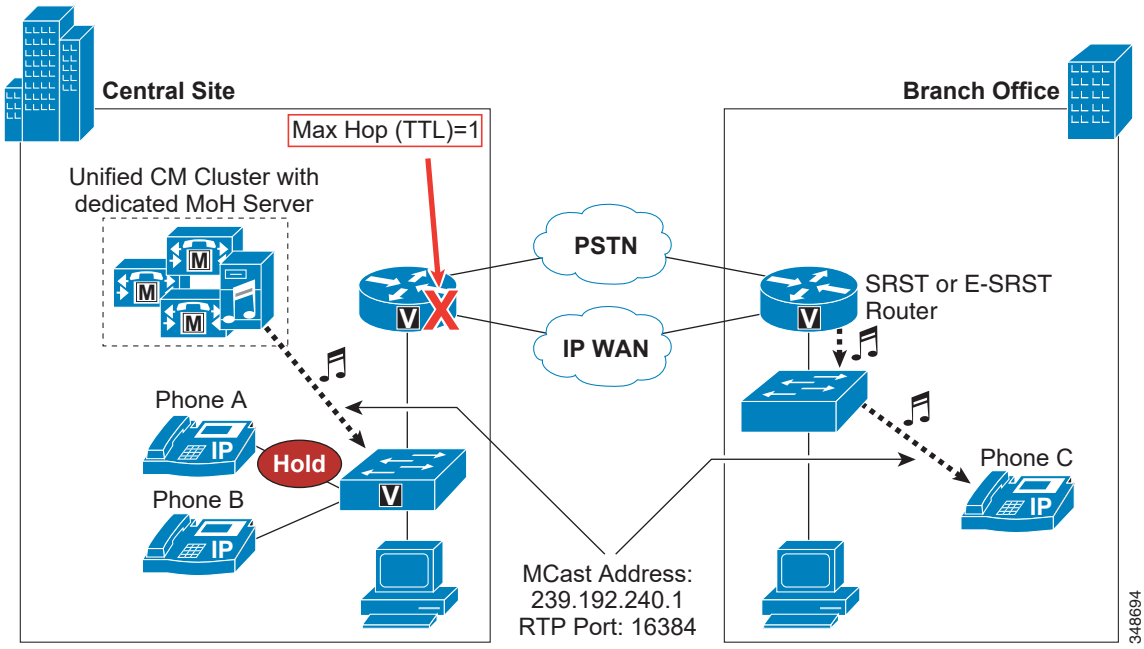
  Configure an ACL on the central-site WAN interface to disallow packets destined to the multicast group address(es) from being sent out the interface.

- Disable multicast routing on the WAN interface

  Do not configure multicast routing on the WAN interface, thus ensuring that multicast streams are not forwarded into the WAN.

Figure 7-14 illustrates streaming multicast MoH from a branch router when it is not in fallback mode. After phone A places phone C on hold, phone C receives multicast MoH from the local SRST router. In this example, the MoH server is streaming a multicast audio source to 239.192.240.1 (on RTP port 16384); however, this stream has been limited to a maximum hop of one (1) to ensure that it will not travel off the local MoH server's subnet and across the WAN. At the same time, the branch office SRST router/gateway is multicasting an audio stream from either flash or a live feed. This stream is also using 239.192.240.1 as its multicast address and 16384 as the RTP port number. When phone A presses the Hold softkey, phone C receives the MoH audio stream sourced by the SRST router.

Figure 7-14 Multicast MoH from Branch Router



When using this method for delivering multicast MoH, configure all devices within the Unified CM cluster to use the same user hold and network hold audio source and configure all branch routers with the same multicast group address and port number. Because the user or network hold audio source of the holder is used to determine the audio source, if you configure more than one user or network hold audio source within the cluster, there is no way to guarantee that a remote holdee will always receive the local MoH stream. For example, suppose a central-site phone is configured with an audio source that uses group address 239.192.254.1 as its user and network hold audio source. If this phone places a remote device on hold, the remote device will attempt to join 239.192.254.1 even if the local router flash MoH stream is sending to multicast group address 239.192.240.1. If instead all devices in the network are configured to use the user/network hold audio source with multicast group address 239.192.240.1 and all branch routers are configured to multicast from flash on 239.192.240.1, then every remote device will receive the MoH from its local router.

Fallback Mode

During fallback mode (when the WAN is down and SRST is active), the branch SRST router can stream multicast MoH to all analog and digital ports within the chassis, thereby providing MoH to analog phones and PSTN callers.

The branch router's configuration for fallback mode multicast MoH is the same as the normal operation configuration. However, which multicast address you configure on the router depends on the intended operation. If you want the branch router to provide multicast MoH to devices only in fallback mode (for example, if MoH received by remote devices is to be sourced from the central-site MoH server during non-fallback mode), then the multicast address and port number configured on the SRST router should not overlap with any of the central-site MoH server audio sources. Otherwise, remote devices might continue to receive MoH from the local router flash, depending on the configured user/network hold audio sources.

Note that, once the branch SRST/gateway router is configured to provide multicast MoH, the router will continue to multicast the MoH stream even when not in fallback mode.

It is also possible to configure the fallback mode to use Cisco Unified Communications Manager Express (Unified CME) in SRST mode, referred to as Enhanced SRST (E-SRST). Fallback mode behavior is still the same, but the configuration commands are slightly different. SRST commands are entered under the Cisco IOS **call-manager-fallback** construct, while the commands for E-SRST mode are entered under **telephony-service**.

There are four methods of providing multicast MoH via SRST:

- SRST multicast MoH from branch router flash
- SRST multicast MoH from a live feed
- E-SRST mode with multicast MoH from branch router flash
- E-SRST mode with multicast MoH from a live feed

For more details on configuration of Cisco Unified SRST and E-SRST, refer to the following documentation:

- *Cisco Unified SRST System Administrator Guide*, available at

  https://www.cisco.com/en/US/products/sw/voicesw/ps2169/products_installation_and_configuration_guides_list.html

- *Cisco Unified Communications Manager Express System Administrator Guide*, available at

  https://www.cisco.com/en/US/products/sw/voicesw/ps4625/products_installation_and_configuration_guides_list.html

- For more information on using Cisco Unified SRST as a multicast MoH resource, refer to the section on *Integrating Cisco Unified Communications Manager and Cisco Unified SRST to Use Cisco Unified SRST as a Multicast MoH Resource* in the latest version of the *Cisco Unified SCCP and SIP SRST System Administrator Guide*, available at

  https://www.cisco.com/en/US/products/sw/voicesw/ps2169/products_installation_and_configuration_guides_list.html

Distributed Multisite Deployments

Multisite IP telephony deployments with distributed call processing typically contain WAN or MAN connections between the sites. These lower-speed links usually cause bandwidth and throughput bottlenecks. To minimize bandwidth consumption on these links, Cisco recommends use of the G.729 codec for all MoH audio streams traversing them. Because the G.729 codec is optimized for voice and not music applications, you should use G.729 only across the WAN/MAN links, where the bandwidth savings far outweighs the lower quality afforded by G.729 for MoH transport.

Unlike with centralized multisite deployments, in situations where G.711 might be required for MoH audio streams traveling across a WAN, MoH audio streams cannot be forced to G.711 in a distributed multisite deployment. Even when MoH servers are placed in a separate Unified CM region and the G.711 codec is configured between this region and the intercluster or SIP trunk's region, the codec of the original voice call is maintained when a call between the two clusters is placed on hold by either phone. Because these intercluster calls are typically encoded using G.729 for bandwidth savings, a MoH stream from either cluster will also be encoded using G.729.

Another option is to provision multicast MoH for intercluster calls across an intercluster trunk (ICT) or SIP trunk. This allows endpoints in one Unified CM cluster to hear multicast MoH streamed from another Unified CM cluster, while making more efficient use of intercluster bandwidth. A properly designed IP Multicast environment is required to take advantage of this feature. For more information on IP Multicast, refer to the documentation available at

https://www.cisco.com/en/US/products/ps6552/products_ios_technology_home.html

Proper multicast address management is another important design consideration in the distributed intercluster environment. All MoH audio source multicast addresses must be unique across all Unified CM clusters in the deployment to prevent possible overlap of streaming resources throughout the distributed network.

Clustering Over the WAN

As its name suggests, clustering-over-the-WAN deployments also contain the same type of lower-speed WAN links as other multisite deployments and therefore are subject to the same requirements for G.729 codec, multicast transport mechanism, and solid QoS for MoH traffic traversing these links.

In addition, you should deploy MoH server resources at each side of the WAN in this type of configuration. In the event of a WAN failure, devices on each side of the WAN will be able to continue to receive MoH audio streams from their locally deployed MoH server. Furthermore, proper MoH redundancy configuration is extremely important. The devices on each side of the WAN should point to an MRGL whose MRG has a priority list of MoH resources with at least one local resource as the highest priority. Additional MoH resources should be configured for this MRG in the event that the primary server becomes unavailable or is unable to process requests. At least one other MoH resource in the list should point to an MoH resource on the remote side of the WAN in the event that resources at the local side of the WAN are unavailable.

Collaboration Endpoints

Revised: March 1, 2018

A variety of endpoints can be used in a Cisco Collaboration deployment. These endpoints range from gateways that support ordinary analog phones in an IP environment to an extensive set of native IP phones offering a range of capabilities.

When deploying endpoints, you need to consider several factors, including authentication, upgrades, signaling protocol, Quality of Service (QoS), and so forth. The collaboration system must be designed appropriately to accommodate these factors.

This chapter summarizes various types of collaboration endpoints and covers design and deployment considerations including high availability and capacity planning. The collaboration endpoints covered in this chapter can be categorized into the following major types:

- Analog Endpoints, page 8-5
- Desk Phones, page 8-8
- Video Endpoints, page 8-14
- Software-Based Endpoints, page 8-22
- Wireless Endpoints, page 8-33
- Mobile Endpoints, page 8-37
- Cisco Virtualization Experience Media Engine, page 8-42
- Third-Party IP Phones, page 8-43

The sections listed above provide information about each endpoint type, including deployment considerations. That information is followed by a discussion related to high availability, capacity planning, and design considerations for effectively deploying endpoints.

Use this chapter to understand the range of available endpoint types and the high-level design considerations that go along with their deployment.

What's New in This Chapter

Table 8-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

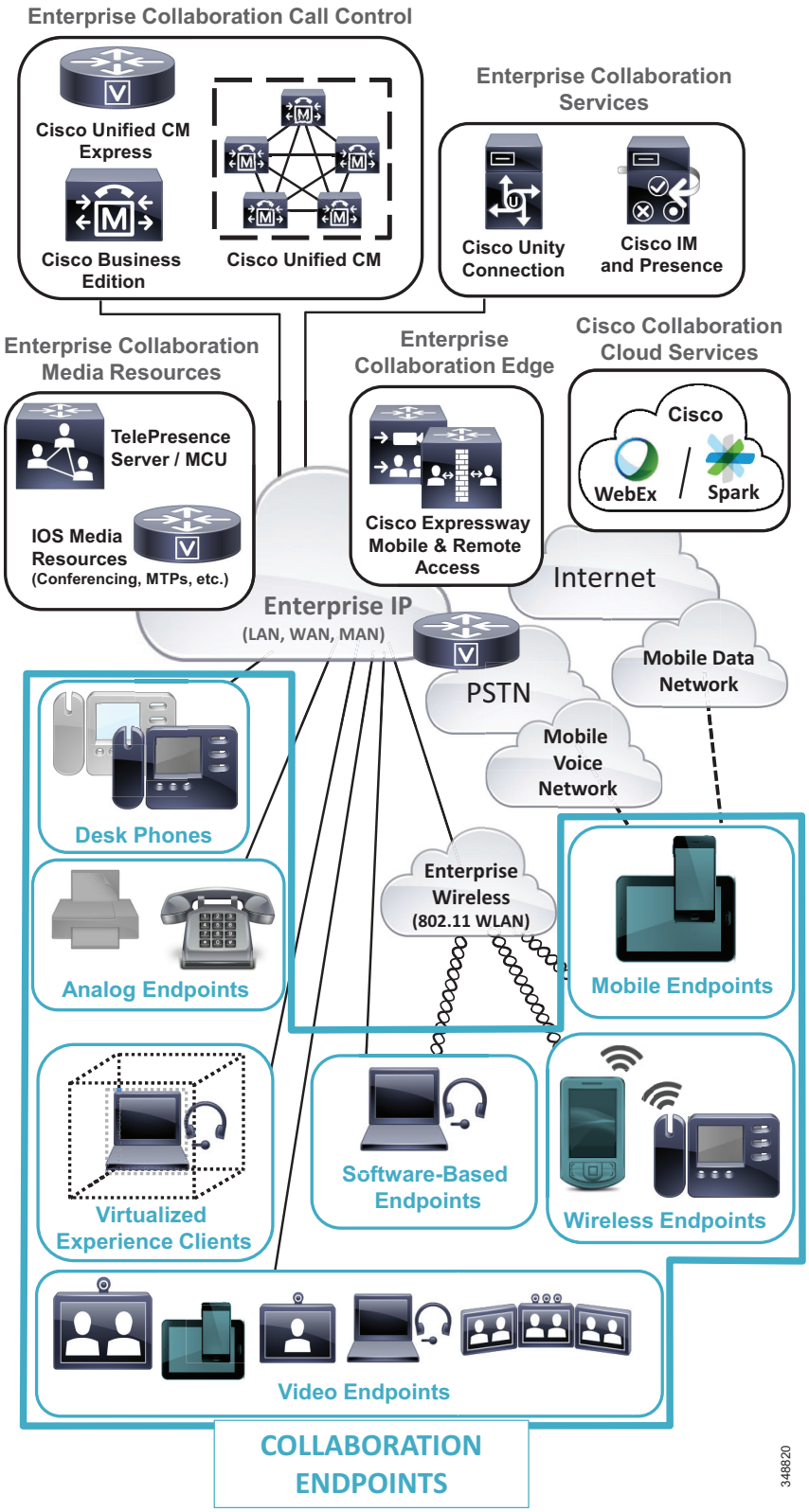
Table 8-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Apple Push Notification service (APNs) for Cisco Jabber and Cisco Spark Apple iOS clients	Deployment Considerations for Mobile Endpoints and Clients, page 8-38	March 1, 2018
OAuth 2.0 with Refresh Token for Cisco Jabber	General Deployment Considerations for Software-Based Endpoints, page 8-29 Deployment Considerations for Mobile Endpoints and Clients, page 8-38	March 1, 2018
Cisco Spark Room Series	Cisco Spark Room Series, page 8-17	March 1, 2018
Cisco IP Phone 8800 Series	Cisco IP Phone 8800 Series, page 8-9	March 1, 2018

Collaboration Endpoints Architecture

Just as there is a variety of endpoint types, as shown in Figure 8-1, there is also a variety of call control, collaboration services, and media resource options that must be considered when deploying collaboration endpoints. Collaboration endpoints rely on enterprise call control and/or cloud-based collaboration for voice and video calling services. Collaboration endpoints also leverage both enterprise on-premises and cloud-based collaboration services such as voice messaging, instant messaging, and presence. Further, these endpoints gain key supplementary services from enterprise media resources such as video and voice conferencing, transcoding, and music on hold.

Figure 8-1 Cisco Collaboration Endpoints Architecture



While there are multiple options when deploying collaboration call control for voice and video services, each call control platform provides endpoint registration, call setup and routing services, and access to provisioned media resources. The high-level call control interactions between endpoints and the enterprise Cisco Unified Communications Manager is described in the following sub-section.

Cisco Unified Communications Manager (Unified CM) Call Control

Call signaling in Cisco Unified Communications Manager (Unified CM), Cisco Business Edition, and Cisco Unified Communications Manager Express (Unified CME) distinguishes between line-side signaling and trunk-side signaling. Whereas trunk-side signaling is used for connecting the entire call processing cluster or router to other servers and gateways, the line side is used for connecting endpoint devices to the call processing platform. The two interfaces are distinct in the services they offer, with the line side offering a rich set of user-oriented features.

Session Initiation Protocol (SIP) and Skinny Client Control Protocol (SCCP) are the two main line-side signaling protocols supported by Cisco call processing platforms. All Cisco endpoints support either or both of these protocols. The set of features supported in both protocols is roughly equivalent, and the choice of which protocol to use is essentially a personal preference in a deployment. However, SIP is the protocol of choice for support of all new features and Cisco endpoints.

Cisco endpoints must be configured with several operating parameters before they can be used to make or receive calls or to run applications. This configuration must be performed in advance on the call processing server or router. Once configured, the call processing platform generates a configuration file for the endpoint to use, and it stores that file on a Trivial File Transfer Protocol (TFTP) server. The endpoints themselves go through a boot-up sequence when powered on. They retrieve this configuration file before they register with the appropriate server, and then they are ready to be used. The endpoints execute the following steps as part of the boot-up sequence:

1. When connected to the access switch, if the endpoint is not plugged in to a power source, it attempts to obtain power from the switch (Power over Ethernet). Wireless and mobile endpoints are not connected to the enterprise network via Ethernet and therefore always derive power from a battery or power outlet.
2. Once power is obtained, if device security is enabled, the endpoint presents its credentials to the security server or network authentication infrastructure.
3. If it is allowed to use the network, the endpoint obtains its network parameters such as IP address, Domain Name Service (DNS) servers, gateway address, and so forth, either through static provisioning in the endpoint or through Dynamic Host Control Protocol (DHCP).
4. The endpoint also obtains a TFTP server address either through static provisioning or through DHCP options.
5. The endpoint then uses the TFTP server address to obtain its configuration files that, among other parameters, details the call processing server(s) or router(s) that the endpoint may associate and register with, the directory numbers that the endpoint must support, and so forth.
6. The endpoint registers with the call processing platform and is available for use.

To confirm which endpoints support registration to Cisco Unified CM, refer to the endpoint data sheets listed in various other sections of this chapter.

Collaboration Endpoint Section 508 Conformance

Regardless of the call control platform you choose, when selecting endpoints and designing your Cisco Collaboration network you should strive to make the telephony features more accessible to users with disabilities, in conformance with Section 255 of the Telecommunications Act and U.S. Section 508.

Observe the following basic design guidelines when configuring your Cisco Unified Communications network to conform to Section 508:

- Enable Quality of Service (QoS) and call admission control on the network to ensure optimal quality of voice and video so that enterprise communications are as clear and precise as possible.
- Configure only the G.711 codec for phones that will be connected to a terminal teletype (TTY) device or a Telephone Device for the Deaf (TDD). Although low bit-rate codecs such as G.729 are acceptable for audio transmissions, they do not work well for TTY/TDD devices if they have an error rate higher than 1% Total Character Error Rate (TCER).
- Configure TTY/TDD devices for G.711 across the WAN, if necessary.
- Enable (turn ON) Echo Cancellation for optimal performance.
- Voice Activity Detection (VAD) does not appear to have an effect on the quality of the TTY/TDD connection, so it may be disabled or enabled. However, Cisco recommends leaving VAD (also known as silence suppression) disabled on Unified CM call control and using the **no vad** command on H.323 and Cisco IOS SIP dial-peers.
- Configure the appropriate *regions* and *device pools* in Unified CM to ensure that the TTY/TDD devices always use G.711 codecs.
- Connect the TTY/TDD to the Cisco Unified Communications network in either of the following ways:
  - Direct connection (Recommended method)

    Plug a TTY/TDD with an RJ-11 analog line option directly into a Cisco FXS port. Any Cisco voice gateway with an FXS port will work. Cisco recommends this method of connection.
  - Acoustic coupling

    Place the IP phone handset into a coupling device on the TTY/TDD. Acoustic coupling is less reliable than an RJ-11 connection because the coupling device is generally more susceptible to transmission errors caused by ambient room noise and other factors.
- If stutter dial tone for audible message waiting indication (AMWI) is required, use an analog phone in conjunction with an FXS port on the Cisco VG Analog Gateways or Analog Telephony Adaptor (ATA). In addition, most Cisco IP Phones support stutter dial tone.
- When you deploy immersive Cisco TelePresence rooms, ensure that ample room is provided to accommodate and provide for unimpeded movement of wheel chairs and other assistive conveyances.

Analog Endpoints

An analog gateway typically is used to connect analog devices such as fax machines, modems, telecommunications device for the deaf (TDD)/teletypewriter (TTY), and analog phones, to the VoIP network so that the analog signal can be packetized and transmitted over the IP network. Analog gateways also provide physical connectivity to the PSTN and other traditional telephony equipment such as PBXs and key systems. Analog gateways include Cisco IOS router-based analog interface or service modules as well as fixed-port standalone gateways. Generally analog gateways rely on Cisco

Unified CM, Cisco Business Edition, Unified CM Express, and even Survivable Remote Site Telephony (SRST) for call control, supplementary services, and in some cases interface registration and configuration. Call control protocols supported across Cisco analog gateways include SIP, H.323, SCCP, and Media Gateway Control Protocol (MGCP).

Standalone Analog Gateways

Cisco standalone analog gateways, including the Cisco Analog Telephony Adaptor (ATA) and Cisco VG Series Gateway, provide connectivity for analog devices such as fax machines, modems, TDD/TTY, paging systems, and analog phones, as well as one or more Ethernet ports for connecting to the IP network. Cisco standalone analog gateways support the FXS analog telephony interface port type only.

For more information on Cisco ATAs, refer to the data sheets and documentation for the ATA 190 Series at:

https://www.cisco.com/c/en/us/products/unified-communications/ata-190-series-analog-telephone-adapters/index.html

For more information on Cisco VG Series Gateways, refer to the data sheets and documentation at:

https://www.cisco.com/c/en/us/products/unified-communications/vg-series-gateways/index.html

Analog Interface Module

Cisco IOS router-based analog interface modules, including network modules (NMs) and voice interface cards (VICs), connect the PSTN and other legacy telephony equipment, including PBXs, analog telephones, fax machines, and key systems, to Cisco multiservice access routers such as the Cisco Integrated Services Router (ISR). Cisco IOS analog interface modules support a wide range of analog telephony interface port types, including FXS, FXO, T1/E1, E&M, and BRI.

Cisco IOS version support is critical for successful deployment of analog interface modules. For more information on Cisco IOS-based analog interface modules, including interface port type and Cisco IOS version support, refer to the data sheets and documentation listed at

https://www.cisco.com/c/en/us/products/collateral/routers/4000-series-integrated-services-routers-isr/datasheet-c78-733646.html

Deployment Considerations for Analog Endpoints

The following sections list important design considerations for deploying analog endpoints.

Analog Connection Types

The choice of analog connection type is typically dictated by the type of analog connection being made. For example, an FXS or E&M interface provides ring and dial tone for basic telephone handsets, while FXO interfaces are used for trunk or tie line connections to a PSTN or to an enterprise PBX. In all cases these interfaces indicate on-hook or off-hook status and the seizure of telephone lines.

With FXO and FXS analog connections there are two types of access signaling methods: loop start or ground start. The type of signaling used is ultimately determined by the type of service from the PSTN. Typically standard telephone land lines use loop start, but business telephone lines and trunks usually rely on ground start. A loop start line does not maintain any current on the circuit until it is in use, whereas a ground start line maintains some current on the line. The use of constant current on the ground

start line typically requires specialized equipment on the PSTN side, which typically makes these lines more expensive than loop start lines. However, with ground start lines, because a loss of current on the line is immediately detected on the far side of the analog connection, the gateway or PBX gets immediate indication regarding connects and disconnects, thus providing better control over the connection. In addition, a ground start trunk reduces the possibility of "glare," or the collision of simultaneous incoming and outgoing calls on the line.

E&M interfaces support different signaling methods, including wink start and immediate start. Wink start is the most common form of E&M signaling, and it relies on a "wink" sequence (on-hook, off-hook, on-hook) indication from the far end in response to an initial off-hook indication at the call origination side before digits can be sent over the interface. In contrast, immediate start signaling relies on a brief pause rather than a response from the far end after the initial off-hook indication before digits are sent.

The analog interface type used in a particular deployment will ultimately be dictated by the interface supported by the PSTN provider or by the equipment deployed in the case of internal analog connections. In all cases, you should use the supported method of signaling for the analog connection type that provides the most visibility and control of the line. For example, with FXS or FXO, ground start is preferred over loop start because of the end-to-end line current which, when broken, can be detected immediately. Likewise, with E&M, wink start is preferred over immediate start because of the positive indication from the far end that digits can be sent.

For additional information on Cisco analog telephony signaling, refer to the documentation available at

https://www.cisco.com/c/en/us/tech/voice/telephony-signaling/index.html

Paging Systems

In some IP telephony deployments, the enterprise IP PBX is integrated with a paging system that allows users to call an extension on the system that forwards the audio broadcast to overhead loudspeakers. These overhead paging systems are useful in workshops, parking lots, and open plant areas where a called party is not near a telephone handset. Integration to these paging systems is done using an analog interface module port.

Cisco analog gateways and interface modules support all traditional analog port types used for paging system integration, including FXO, FXS, and E&M. When integrating with overhead paging systems, ensure that the appropriate analog interface module port type, signaling, and configuration are used as required by the paging system being integrated. The port type, signaling, and configuration will ultimately be dictated by the paging system.

An example of an E&M interface integration to an overhead paging system is available at

https://www.cisco.com/c/en/us/support/docs/voice/analog-signaling-e-m-did-fxs-fxo/27627-e-mpaging.html

Quality of Service

When configuring network-level quality of service (QoS), Cisco analog gateways such as the standalone Cisco VG Series and the Cisco IOS-based analog interface modules can be trusted and their packet markings honored. By default they mark their voice media and signaling packets with appropriate Layer 3 values (voice media as DSCP 46 or PHB EF; call signaling as DSCP 24 or PHB CS3), which match Cisco QoS recommendations for appropriate voice media and signaling marking, so as to ensure end-to-end voice quality on a converged network.

Desk Phones

The Cisco IP Phone portfolio includes the following family of desk phones:

- Cisco Unified IP Phone 7900 Series, page 8-8
- Cisco IP Phone 8800 Series, page 8-9
- Cisco Unified SIP Phone 3900 Series, page 8-10
- Cisco DX Series, page 8-10

Cisco Unified IP Phone 7900 Series

The Cisco Unified IP Phone 7900 Series of endpoints consists of several models and feature sets. In general, all phones in the Unified IP Phone 7900 Series provide the same basic set of enterprise IP telephony features such as call hold, call transfer, call forwarding, and so forth. However, the 7900 Series also provides features and functions well beyond the traditional enterprise IP telephony feature set, including support for IP-based phone services to enable presence, messaging, mobility, security, and other network-based applications and services. The Cisco Unified IP 7900 Series supports both SCCP and SIP signaling protocols for registering and communicating with the Cisco call processing platforms.

In some cases additional line keys can be added to Unified IP Phone 7900 Series devices by physically attaching a key expansion module such the Cisco Unified IP Phone Expansion Module 7916. This gives administrative assistants and other users the ability to answer and/or determine the status of a number of lines beyond the current line capability of their desk phone. Some Unified IP Phone 7900 Series models are capable of supporting up to two Cisco Unified IP Phone Expansion Modules, but the use of an external power adaptor may be required.

Note When two Expansion Modules are used with a single phone, the second module must be the same model as the first one.

For more information about the Cisco Unified IP Phone 7900 Series, refer to the data sheets and documentation at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/unified-ip-phone-7900-series/index.html

Cisco IP Phone 7800 Series

The Cisco IP Phone 7800 Series of endpoints includes a number of models ranging from the single-line Cisco IP Phone 7811 to the larger, more advanced 16-line Cisco IP Phone 7861. These phone models have LCD displays, built-in speakerphone, and PC ports. In general, all of the phones in the IP Phone 7800 Series provide enterprise IP telephony features such as hold, call transfer, call forwarding, and so forth. The Cisco IP 7800 Series supports SIP signaling protocol for registering and communicating with the Cisco call processing platforms.

Note Starting with Cisco IP Phone 7800 Series firmware version 11.0(1)and with Cisco Expressway X8.7 and later versions, the 7800 Series officially supports Cisco Expressway as an alternative to VPN access. Expressway provides enterprise firewall traversal for 7800 Series voice calls.

For more information about the Cisco IP Phone 7800 Series, refer to the data sheets and product documentation available at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/unified-ip-phone-7800-series/index.html

Cisco IP Phone 8800 Series

The Cisco IP Phone 8800 Series of endpoints delivers a highly secure and comprehensive feature set with support for wideband audio. For example, the new Cisco IP Conference Phone 8832 provides dynamic, detailed sound with low distortion and low-frequency clarity. Both wired and Digital Equipment Cordless Telephony (DECT) wireless extension microphones are available for 360-degree coverage in conference room deployments. In addition, the IP Phone 8800 Series personal endpoints, from the 8811 to the 8865, provide a range of features. Some models in this series, such as the 8845, 8851, 8861, and 8865, provide support for Bluetooth and Intelligent Proximity for Mobile Voice as well as smartphone or tablet charging via on-board USB ports. New Key Expansion Modules (KEM) have been introduced for the 8800 Series that provide dual LCD support to maximize the viewing area and improve the user experience. A new audio KEM and video KEM have been introduced. Also, the 8845 and 8865 provide HD 720p built-in video camera support. In general, all of the phones in the IP Phone 8800 Series provide an identical set of enterprise IP telephony features such as call hold, call transfer, call forwarding, and so forth. These endpoints support SIP signaling protocol for registering and communicating with the Cisco call processing platforms.

Note Starting with Cisco IP Phone 8800 Series firmware version 11.0(1) and with Cisco Expressway X8.7 and later versions, the 8800 Series phone models officially support Cisco Expressway as an alternative to VPN access. Expressway provides enterprise firewall traversal for 8800 Series voice and video calls.

Note Starting with Cisco IP Phone 8800 Series firmware version 11.5, the 8800 Series phone models support Enhanced Line Mode, which allows for the assignment of programmable lines or features such as speed dials to all 10 line keys. Prior to this firmware enhancement, only 5 programmable line keys could be used on the phone. In addition, firmware version 12.0 adds support to Enhanced Line Mode for call park, extension mobility cross cluster, group pickup, and hunt groups.

For more information about the Cisco IP Phone 8800 Series, refer to the data sheets and other documentation at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/unified-ip-phone-8800-series/index.html

Cisco Unified SIP Phone 3900 Series

The Cisco Unified SIP Phone 3900 Series provides cost-effective, entry-level endpoints that support a single line and provide a basic set of enterprise IP telephony capabilities and basic supplementary features such as mute, call hold, and call transfer. The Cisco Unified SIP Phone 3900 Series has a two-line liquid crystal display (LCD) screen and a half-duplex or full-duplex speakerphone (depending on the model). The Cisco Unified SIP Phone 3900 Series supports the SIP signaling protocols for registering and communicating with Cisco call processing platforms.

Note The Cisco Unified SIP Phone 3900 Series does not support features such as CTI (for Jabber phone control), speed dials, or Built-in Bridge for Silent Monitoring and Recording. The Cisco IP Phone 7800 Series or 8800 Series are recommended for environments requiring the full set of enterprise grade IP telephony features.

For more information about the Cisco Unified SIP Phone 3900 Series, refer to the data sheets and documentation at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/unified-sip-phone-3900-series/index.html

Cisco DX Series

The DX Series of endpoints delivers integrated Unified Communications, high-definition (HD) video, and collaboration applications and services. The Cisco DX Series endpoints provide wideband audio and HD video for enterprise-class communications with integrated 7 to 23 inch (model dependent) multi-touch LCD display and front-facing camera. These devices run secure Android operating system and provide access to a variety of integrated collaboration and communication applications including calendaring, corporate directory searches, email, Jabber IM and presence, visual voicemail, and WebEx conferencing as well as AnyConnect VPN for secure network attachment. In addition, as an open Android platform, these devices are capable of accessing the Google Play store for access to many third-party applications that enable additional features and functionality. These endpoints also provide a variety of external interfaces for attaching accessories, including: HDMI for connecting external devices such as a laptop or external display (model dependent); USB for keyboard, mouse, or wired headset attachments; and Bluetooth for connecting a wireless headset, keyboard, and/or mouse, or for leveraging Intelligent Proximity for Mobile Voice.

The DX Series endpoints support SIP signaling protocol for registering and communicating with Cisco call processing platforms. Cisco Unified Communications Manager is required to deploy and support the DX Series.

Note Starting with Cisco DX Series firmware version 10.2.4, the DX Series supports Cisco Expressway as an alternative to VPN access. Expressway provides enterprise firewall traversal for DX Series voice and video, as well as the built-in Jabber IM application.

For more information about the Cisco DX Series, refer to the data sheets and documentation at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/desktop-collaboration-experience-dx600-series/index.html

Deployment Considerations for Cisco Desk Phones

The following sections list important design considerations when deploying Cisco desk phones.

Firmware Upgrades

Most commonly, and by default, IP phones upgrade their images using HTTP, which uses Port 6970, from TFTP services integrated into one or more of the call processing platforms. When HTTP is not available, IP phones use TFTP, which is a UDP-based protocol from the same TFTP services. With this arrangement, all the phones obtain their images directly from these TFTP services. This method works well for a relatively small number of phones or if all of the phones are located in a single campus region that has a LAN environment with essentially unlimited bandwidth.

For larger deployments that use centralized call processing, upgrading phones in branch offices that are connected to the central data center by low-speed WAN links, can require a large amount of data traffic over the WAN. The same set of files will have to traverse the WAN multiple times, once for each phone. Transferring this amount of data is not only wasteful of the WAN bandwidth but can also take a long time as each data transfer competes with the others for bandwidth. Moreover, due to the nature of TFTP protocol, some phones might be forced to abort their upgrades and fall back to the existing version of the code.

Note During the upgrade, the Cisco IP Phones 7800, 8800, and DX Series stay in service, unlike the 7900 Series phones. The 7800, 8800, and DX Series phones download and store the new firmware in their memory while still maintaining their active status, and they reboot with the new firmware only after a successful download.

Two methods are available to alleviate problems created by the need to upgrade phones over the WAN. One method is to use a local TFTP server just for the upgrades. The administrator can place a TFTP server in branch offices (particularly in branches that have a larger number of phones, or whose WAN link is not speedy or robust), and can configure the phones in those offices to use that particular TFTP server just for new firmware. With this change, phones will retrieve new firmware locally. This upgrade method would require the administrator to pre-load the phone firmware on the TFTP server in the branch and manually configure the TFTP server address in the **load server** parameter in the affected phone configurations. Note that the branch router may be used as a TFTP server.

The second method to upgrade phones without using the WAN resources excessively is to use the Peer File Sharing (PFS) feature. With this feature, typically only one phone of each model in the branch downloads each new firmware file from the central TFTP server. Once the phone downloads the firmware file, it distributes that file to other phones in the branch. This method avoids the manual loading and configuration required for the load server method.

The PFS feature works when the same phone models in the same branch subnet arrange themselves in a hierarchy (chain) when asked to upgrade. They do this by exchanging messages between themselves and selecting the "root" phone that will actually perform the download. The root phone sends the firmware file to the second phone in the chain using a TCP connection; the second phone sends the firmware file to the third phone in the chain, and so on until all of the phones in the chain are upgraded. Note that the root phone may be different for different files that make up the complete phone firmware.

Power Over Ethernet

Deploying desk phones with inline power-capable switches enables these endpoints to derive power over the Ethernet network connection, thus eliminating the need for an external power supply as well as a wall power outlet. Inline power-capable switches with uninterruptible power supplies (UPS) ensures that power over Ethernet (PoE) capable IP desk phones continue to receive power during power failure situations. Provided the rest of the telephony network is available during these periods of power failure, then IP phones should be able to continue making and receiving calls.

Depending on the type of desk phone and the PoE standard supported by both the desk phone and the inline power-capable switch, in some cases the power budget of the inline powered switch port may be exceeded. This typically occurs when attaching key extension modules or other power consuming attachments such as USB cameras. In these situations, the phone may need to be powered using a wall outlet and external power supply or else the switch providing the power may need to be upgraded.

Note In addition to using the inline power from the access switch or local wall power, a Cisco Unified IP Phone can also be supplied power by a Cisco Unified IP Phone power injector. The Cisco Unified IP Phone power injector connects Cisco Unified IP Phones to Cisco switches that do not support inline power or to non-Cisco switches. The Cisco Unified IP Phone power injector is compatible with most Cisco Unified IP Phones. It has two 10/100/1000 Base-T Ethernet ports. One Ethernet port connects to the switch access port and the other connects to the Cisco Unified IP Phone.

Quality of Service

When configuring network-level quality of service (QoS), Cisco desk phones such as the Cisco Unified IP Phone 7900, 8800, and DX Series can be trusted and their packet markings honored. By default these endpoints mark their voice media and signaling packets with appropriate Layer 3 values (voice media as DSCP 46 or PHB EF; call signaling as DSCP 24 or PHB CS3), which match Cisco QoS recommendations for appropriate voice media and signaling marking, to ensure end-to-end voice quality on a converged network. While many Cisco desk phones support the attachment of a desktop computer, Cisco desk phones are capable of separating the voice and data traffic, placing voice traffic onto the voice VLAN and data traffic from the desktop onto the data VLAN. This enables the network to extend trust to the phone but not to the PC port of the phone. However, for multipurpose devices such as the Cisco DX Series endpoints, which are capable of generating both voice and data traffic without an attached desktop computer, both voice and data traffic will traverse the same VLAN. In these cases, whether the device is attached to the voice or data VLAN, extending trust to these devices might not be advisable. Instead, re-marking the traffic based on port and protocol will ensure that all traffic is appropriately marked regardless of the VLAN it traverses.

In deployment where there are concerns about the potential volume of data traffic generated by multipurpose devices such as the Cisco DX Series and the possibility of adversely impacting real-time voice and video traffic, these devices should be deployed in the data VLAN or in a separate VLAN. This will alleviate concerns about impacting call quality of voice and video-only devices. Further, with packet re-marking based on ports and protocols, priority treatment can still be provided within the VLAN to real-time traffic generated by these multipurpose devices.

Note While many Cisco desk phones support Link Layer Discovery Protocol for Media Endpoint Devices (LLDP-MED), they do so only for VLAN and Power over Ethernet negotiation. Cisco Unified IP Phones do not honor DSCP and CoS markings provided by LLDP-MED.

SRST and Enhanced SRST

When deploying Cisco desk phones in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. By leveraging Survivable Remote Site Telephony (SRST) or Enhanced SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for the desk phones when connectivity to the centralized call processing platform is lost. However, the set of available user-facing features is much smaller when a device is registered to SRST than when the phone is registered to Unified CM.

Secure Remote Enterprise Attachment

Cisco desk phones can be securely connected to the enterprise network from remote locations using VPN or VPN-less solutions.

In the case of VPN-based connectivity, desk phones can be located behind a VPN router that creates a secure VPN tunnel to the Cisco Adaptive Security Appliance (ASA) or other VPN head-end concentrator at the enterprise edge. Alternatively, some phone models support a native built-in VPN client that provides VPN connectivity within the phone itself for voice traffic (media and signaling) of the device, but not for the PC or data traffic. In this case the phone creates a secure VPN tunnel to the Cisco ASA within the enterprise. The native built-in VPN client is supported only on certain phone models, including the Cisco Unified IP Phone 7945, 7965, and 7975, as well as the 8800 Series phones. For more information on built-in VPN on Cisco Unified IP Phones, refer to the latest version of the *Security Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

Note The 7800 Series does not support built-in VPN. The 7800 Series does support Expressway Mobile and Remote Access starting with firmware version 11.0(1) and with Expressway X8.7 and later versions.

Cisco IP Phone 7800 Series, IP Phone 8800 Series, and DX Series endpoints are able to leverage mobile and remote access functionality of the Cisco Expressway solution. This firewall traversal solution relies on TLS reverse proxy connectivity to the enterprise, as provided by the Cisco Expressway-E and Expressway-C servers for registration to Unified CM call control for voice and video calling. For more information about mobile and remote access capabilities of the Cisco Expressway solution, refer to the solution information and product documentation available at

https://www.cisco.com/c/en/us/products/unified-communications/expressway-series/index.html

Intelligent Proximity

Intelligent Proximity refers to features that leverage proximity-based connections between Cisco hardware endpoints and mobile devices.

Intelligent Proximity for Mobile Voice capabilities available on the Cisco DX Series and select 8800 Series endpoints rely on the use of Bluetooth pairing between the DX or 8800 endpoint and a cellular or smart phone.

Bluetooth paired mobile devices are able to invoke two features:

- Hands-free audio, providing the ability to send audio of a cellular terminated call through the DX Series, 8845, 8851, 8861, or 8865 IP endpoint speaker or handset. Audio play-out of the cellular terminated call can be moved back and forth between the DX, 8845, 8851, 8861, or 8865 and the mobile device. In addition, because the Bluetooth paired mobile device appears on the 8845, 8851, 8865, or DX Series endpoints as another line, cellular calls on the Bluetooth paired mobile device can be initiated using the DX or 8800 IP endpoint.
- Mobile contact and call history sharing, providing the ability to share mobile device contact and call history sharing with the DX Series, 8845, 8851, 8861, or 8865 endpoints.

Because Intelligent Proximity for Mobile Voice relies on Bluetooth pairing, there is no requirement to run an application or client on the mobile device. All communication and interaction occurs over the standard-based Bluetooth interfaces.

The Intelligent Proximity for Mobile Voice feature set on the DX Series endpoints and the 8845, 8851, 8861, and 8865 IP phones is compatible with the Unified Mobility feature set, including single number reach (SNR), remote destination and desk phone pickup, two-stage enterprise dialing, and mobile voicemail avoidance. In the case of the 8845, 8851, 8861, and 8865 IP phones, Intelligent Proximity for Mobile Voice is compatible with Cisco Jabber mobile clients. When a Jabber client running on a mobile device is paired with the 8845, 8851, 8861, or 8865 IP phone, the audio portion of a Jabber call may be played out using the 8845, 8851, 8861, or 8865 handset or speaker while the video portion of the call continues to play on the Jabber mobile client. In the case of DX Series endpoints, Intelligent Proximity for Mobile Voice functionality is limited exclusively to the cellular line of Bluetooth paired mobile devices running Jabber.

The Intelligent Proximity for Mobile Voice feature set on the DX Series and select 8800 Series endpoints requires firmware version 10.1.1 or later.

Note Starting with Cisco IP Phone 8800 Series firmware version 11.0(1), the 8845, 8851, 8861, and 8865 phones support Cisco Unified CM Application Dial Rules that allow for dialing imported smartphone contacts over a Cisco VoIP network.

For more information about Intelligent Proximity for Mobile Voice, refer to the product documentation for the Cisco DX Series and 8800 Series endpoints as well as the information at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/intelligent-proximity.html

Video Endpoints

Cisco video endpoints provide IP video telephony features and functions similar to IP voice telephony, enabling users to make point to point and point to multi-point video calls. Cisco offers the following video-capable endpoints:

- Cisco Jabber software-based desktop clients such as Cisco Jabber for Windows
- Cisco Unified IP Phones 8800 Series (8845 or 8865) with built-in camera
- Cisco DX Series with built-in camera
- Cisco TelePresence System EX, MX, SX, and IX Series
- Cisco Spark Room Kit Series

Cisco video endpoints deliver high-quality video for all user types and environments within any organization. Cisco video endpoints are classified into families based on the features they support, hardware screen size, and environment where the endpoint is deployed. This section categorizes the Cisco video endpoint families into personal, multipurpose, and immersive endpoints groups.

Personal Video Endpoints

Personal video endpoints provide a high-quality, face-to-face video calling experience for personal workspaces.

Cisco Jabber Desktop Video

Cisco Jabber software-based desktop clients, such as Cisco Jabber for Windows, are able to send and receive video when running on a desktop computer with an integrated or USB attached camera. These video-capable software-based endpoints register and communicate with Unified CM call control and operate as a SIP single-line voice and video enabled phone. These endpoints support the primary and backup registration redundancy mechanisms as provided by Unified CM. The Cisco Jabber software-based endpoint processes video on the computer where it is installed. The quality of the decoding and encoding depends on the availability of CPU and memory resources on that computer.

For additional information on Cisco Jabber desktop clients, refer to Software-Based Endpoints, page 8-22.

For more information about the video capabilities of Cisco Jabber for Windows, refer to the data sheet and product documentation available at

https://www.cisco.com/c/en/us/products/unified-communications/jabber-windows/index.html

Cisco IP Phone 8800 Series

Some models in this series, specifically the 8845 and 8865, provide support for HD 720p with an integrated video camera. The main difference between the 8845 and the 8865 is that the 8865 also provides support for smartphone or tablet charging via on-board USB ports. Also, the 8865 provides support for up to three Key Expansion Modules. In general, the 8845 and 8865 provide an integrated video experience along with enterprise IP telephony features such as call hold, call transfer, call forwarding, and so forth. These endpoints support SIP signaling protocol for registering and communicating with the Cisco call processing platforms.

For more information about the Cisco IP Phone 8800 Series, refer to the data sheets and documentation at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/unified-ip-phone-8800-series/index.html

Cisco DX Series

The Cisco DX Series endpoints are capable of transmitting video by means of the built-in front-facing camera. These endpoints are capable of receiving and displaying video natively on their screens with a variety of video resolutions and frame rates. The video capabilities of these phones can be enabled and disabled or tuned as desired from the Cisco call control platform configuration pages.

These devices register and communicate with Unified CM using SIP signaling protocol.

For more information about the Cisco DX Series video capabilities, refer to the data sheets and product documentation available at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/desktop-collaboration-experience-dx600-series/index.html

Cisco TelePresence System EX90

The Cisco TelePresence System EX90 video endpoint takes the personal desktop solution to a next level of experience with support for full high definition (HD) video calls and added features such as content sharing. The EX90 has a wide screen with support for the multisite feature that provides the ability to add participants into a Cisco TelePresence call and dual display for content sharing.

The Cisco TelePresence System EX90 video endpoint registers and communicates with Unified CM by means of the SIP signaling protocol.

For more information about the Cisco TelePresence EX90 video endpoint, refer to the data sheets and product documentation available at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/telepresence-system-ex-series/index.html

Multipurpose Video Endpoints

Multipurpose video endpoints enable any size meeting room to become a telepresence room by providing high quality point-to-point or multipoint video collaboration with content sharing.

Cisco TelePresence System MX Series

The MX Series of Cisco TelePresence endpoints provide highly integrated collaboration room systems that are classified as multipurpose room systems. These video endpoints are simple to use and easy to install, providing video calling and content sharing during presentations. They are cost-effective endpoints that can transform any room or existing meeting space into a multipurpose conference room providing full high definition (HD) video calling. There are four variants of the MX Series:

- MX800 single or dual 70-inch screen integrated TelePresence system
- MX700 is a dual 55-inch screen integrated TelePresence system
- MX300 G2 is a 55-inch screen integrated TelePresence system
- MX200 G2 is a 42-inch screen integrated TelePresence system

These endpoints register to Unified CM using the SIP signaling protocol.

For more information about the Cisco TelePresence System MX Series of video endpoints, refer to the data sheets and product documentation available at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/telepresence-mx-series/index.html

Cisco TelePresence SX Series

The Cisco TelePresence SX Series endpoints are flexible integrators that can turn any flat-panel display into a powerful Cisco TelePresence system. The SX Series video endpoints are designed for HD video and multiparty conferencing, with the flexibility to accommodate various room sizes. This is an ideal solution for small to mid-size business and enterprises looking for a cost effective TelePresence-enabled conference room solution. The SX series video endpoints provide the following options:

- SX10 is an all-in-one system codec with integrated camera.
- SX20 is a codec with one of three camera options, and it supports the multisite features, providing the ability to add up to three more participants in a Cisco TelePresence call.
- SX80 is a codec that includes integrator packages supporting different camera and touch panel options.

These endpoints register to Unified CM using the SIP signaling protocol.

For more information about the Cisco TelePresence SX Series video endpoint, refer to the data sheets and product documentation available at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/telepresence-quick-set-series/index.html

Cisco Spark Room Series

The Cisco Spark Room Series endpoints include both highly integrated collaboration room systems that can transform any room or existing meeting space into a multipurpose conference room as well as flexible integrators that can turn any flat-panel display into a powerful Cisco TelePresence system. The Cisco Spark Room Series video endpoints are designed for 4K ultra HD video and screen sharing and multi-party conferencing, with the flexibility to accommodate various room sizes.

There are four variants of the Cisco Spark Room Series:

- Cisco Spark Room Kit is an all-in-one system codec with integrated camera
- Cisco Spark Room Kit Plus is a quad camera system with separate codec
- Cisco Spark Room 55 is a 55-inch screen integrated TelePresence system
- Cisco Spark Room 70 is a single or dual 70-inch screen integrated TelePresence system

These endpoints register to Cisco Unified CM using the SIP signaling protocol or to the Cisco Collaboration Cloud using HTTPS.

For more information about the Cisco Spark Room Series video endpoints, refer to the data sheets and product documentation available at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/spark-room-series/index.html

Immersive Video Endpoints

Immersive video endpoints enable the best possible in-person telepresence video collaboration experience, where attendees across multiple locations feel as though they are in the same room.

Cisco TelePresence IX5000 Series

The Cisco TelePresence IX5000 Series raises the standard for "in-person" collaboration with an industry first H.265 three-screen TelePresence system. This immersive system is not only easy to use but also very easy to set up. The system comes in two variants: the single-row 6-seat IX5000 system and the dual-row 18-seat IX5200 system. These systems are capable of delivering three simultaneous high-definition (1080p, 60 fps) video streams and two content sharing streams (1080p, 30 fps). These endpoints register to Unified CM using the SIP signaling protocol.

For more information about the Cisco TelePresence IX5000 Series immersive video systems, refer to the data sheets and product documentation available at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/ix5000-series/index.html

General Deployment Considerations for Video Endpoints

The following sections list important design considerations for deploying video endpoints.

Quality of Service

When configuring network-level quality of service (QoS), Cisco video endpoints (including Cisco DX Series and Cisco TelePresence System devices) generally mark traffic at Layer 3 according to Cisco general QoS guidelines related to voice and video packet marking (voice media as DSCP 46 or PHB EF; desktop video media as DSCP 34 or PHB AF41; telepresence video media as DSCP 32 or PHB CS4; call signaling as DSCP 24 or PHB CS3), and therefore these devices can be trusted. In the case of personal desktop video endpoints, including Cisco DX Series devices, both voice and video media packets are marked as DSCP 34 or PHB AF41 to preserve lip synchronization during a video call.

While proper network QoS configuration is essential even when the endpoint marking is trusted, Cisco recommends ensuring that sufficient bandwidth is provisioned on the network and then using network-based policing and rate limiting to ensure that all endpoints do not consume more network bandwidth than they should. Software-based video-capable endpoints do present challenges when they do not or cannot mark traffic appropriately. In these situations, typical guidance is to re-mark media and signaling traffic within the network from best-effort to appropriate and recommended values (voice media as DSCP 46 or PHB EF; desktop video and voice media for video calls as DSCP 34 or PHB AF41; telepresence video media as DSCP 32 or PHB CS4; call signaling as DSCP 24 or PHB CS3) based on protocols and/or port numbers.

In the case of software-based Cisco Jabber for Windows, appropriate Layer 3 DSCP QoS marking can be applied to audio and video streams based on voice and video media source port numbers using Microsoft Windows group policies.

For more information about Cisco Jabber for Windows QoS with Microsoft Windows group policies, refer to the Quality of Service configuration information in the latest version of *On-Premises Deployment for Cisco Jabber*, available at

https://www.cisco.com/c/en/us/support/unified-communications/jabber-windows/products-installation-guides-list.html

Note While some Cisco video-capable endpoints support Link Layer Discovery Protocol for Media Endpoint Devices (LLDP-MED), they do so only for VLAN and Power over Ethernet negotiation. Cisco video endpoints do not honor DSCP and CoS markings provided by LLDP-MED.

For more information on video endpoint network bandwidth consumption and QoS marking and classification, see the section on WAN Quality of Service (QoS), page 3-37.

Inter-VLAN Routing

When deploying video endpoints on networks with voice and data VLAN separation, it is important to consider software-based video-capable endpoints as well as hardware-based video endpoints that need to access resources. Because software-based endpoints running on a desktop computer are primarily attached to the data VLAN, inter-VLAN routing should be configured and allowed so that voice traffic from these endpoints on the data VLAN can reach endpoints on the voice VLAN. Likewise, if hardware-based video endpoints such as the Cisco TelePresence System endpoints need access to network resources such as directory or management services deployed on the data VLAN, inter-VLAN routing must be allowed.

SRST and Enhanced SRST

When deploying video endpoints in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. By deploying SRST or Enhanced SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for most video endpoints when connectivity to the centralized call processing platform is lost. The set of available user-facing features is much smaller when a video endpoint is registered to SRST than when the application is registered to Unified CM. Specifically, video endpoint devices registered to SRST will be capable of making and receiving only voice calls (audio-only). SRST is not supported with the Cisco TelePresence System video endpoints. However, starting with Cisco IOS Release 15.3(3)M using phone load firmware 9.4.1 or later, Enhanced SRST supports making and receiving video calls with some video endpoints during WAN failure. For details on Enhanced SRST video support for various phone models, refer to the Cisco Unified IP Phone documentation available at

https://www.cisco.com/

Secure Remote Enterprise Attachment

Cisco video endpoints can be securely connected to the enterprise network from remote locations using VPN or VPN-less solutions.

In the case of VPN-based connectivity, all video endpoints can be located behind a VPN router that creates a secure VPN tunnel to the Cisco Adaptive Security Appliance (ASA) or other VPN head-end concentrator at the enterprise edge. In addition, the Cisco Unified IP Phone 8800 Series supports a native built-in VPN client, which provides VPN connectivity within the phone itself for voice and video traffic (media and signaling) without the need for a VPN router.

For VPN-less connectivity, Cisco TelePresence endpoints running TC firmware (EX, MX, C, and SX Series endpoints) as well as DX Series endpoints are able to leverage mobile and remote access functionality of the Cisco Expressway solution. This firewall traversal solution relies on TLS reverse proxy connectivity to the enterprise as provided by the Cisco Expressway-E and Expressway-C servers

for registration to Unified CM call control for voice and video calling. For additional information on mobile and remote access capabilities of the Cisco Expressway solution, refer to the solution information and product documentation available at

https://www.cisco.com/c/en/us/products/unified-communications/expressway-series/index.html

Intelligent Proximity

As previously mentioned, Intelligent Proximity refers to features that leverage proximity-based connections between Cisco hardware endpoints and mobile devices.

Intelligent Proximity for Mobile Voice capabilities available on the Cisco IP Phone 8800 Series or the Cisco DX Series rely on Bluetooth pairing between the DX endpoint and a cellular or smart phone, enabling hands-free audio and mobile contact and call history sharing.

For more information on Intelligent Proximity and Bluetooth pairing, see Intelligent Proximity, page 8-13.

Video Interoperability

Video interoperability is the audio and video support for point-to-point calls between Cisco TelePresence System video endpoints, other Cisco Collaboration video endpoints, and third-party video endpoints. Previously, video interoperability between different families of video endpoints was possible only with the insertion of a video component between endpoints, such as a video transcoder or a multipoint control unit (MCU).

Cisco Unified CM not only offers native video interoperability between different video endpoint family types, but also provides better video interoperability in general with H.264 codec negotiation in SIP and H.323 protocols and enable the endpoints to negotiate high definition (HD) resolutions when available. Video interoperability, however, is dependent on the endpoints to support the interoperation.

Video interoperability in Unified CM also enables Cisco TelePresence System video endpoints to communicate with non-video endpoints, provided that the installed firmware supports such interoperability. For further information, refer to the *Cisco TelePresence Interoperability Database*, available at

https://tp-tools-web01.cisco.com/start/

Additionally, Cisco Unified CM provides support for enhanced interoperability with call agents other than Unified CM. Through scripting, Unified CM supports the following features:

- SIP transparency — The ability to pass through known and unknown message components
- SIP normalization — Transformations on inbound and outbound SIP messages and content bodies

The primary motivation for video interoperability support is to facilitate the interaction of a diverse set of video endpoints without the need for deploying an expensive hardware-based DSP infrastructure that would otherwise be required. There are additional benefits that can be derived from the use of advanced conferencing and transcoding resources (for example, active presence where participants of multi-point conferences can see the active speaker); however, the desired feature set and video calling needs will dictate when and where those advanced resources would be required.

The following sections present general considerations and recommendations for the use of video interoperability:

- Video Interoperability Architecture, page 8-21
- Design Considerations for Video Interoperability, page 8-21

Video Interoperability Architecture

The video interoperability architecture includes the following elements:

- Video interoperability support available with Cisco Unified CM
- Two different video endpoint family types (Cisco TelePresence System video endpoints, other Cisco Collaboration video endpoints such as the Cisco DX80, or third-party endpoints) engaged in a video call

The following sections offer further information about the scope of the video interoperability support:

- Video Interoperability Test Cases, page 8-21
- Limitations of Video Interoperability, page 8-21

Video Interoperability Test Cases

In most cases a video endpoint that supports SIP or H.323 without using proprietary signaling would be able to interoperate with a Cisco Collaboration video endpoint that supports video interoperability. For specific information on the scope of the interoperability between common sets of deployed devices and general information about the testing that was conducted to validate these more common examples of interoperability, refer to the Cisco Collaboration Systems documentation available at

https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/uc_system/unified/communications/system/ucstart.html

Limitations of Video Interoperability

While video interoperability support attempts to enable any-to-any point-to-point video call interoperability, it is important to note that not all features of an individual video endpoint can be supported when interoperating with another endpoint. There are many reasons for this. For example, incompatibilities between different call control protocols could render a feature unavailable or offer a different representation of that feature. H.264 video media parameters can be represented differently in H.323 than in SIP, as another example. H.323 also does not have support for presence, but presence is quite commonly supported in SIP. Skinny Client Control Protocol (SCCP) does not have any notion of application sharing, which is commonly available in SIP and H.323 endpoint implementations. For instance, an SCCP user trying to share his/her PC screen would be hampered because Binary Floor Control Protocol (BFCP) and H.239 are not available with SCCP.

Design Considerations for Video Interoperability

The following areas should be considered when implementing the video interoperability capabilities of Unified CM:

- Guideline and Restrictions for Video Interoperability, page 8-22
- Quality of Service (QoS) and Call Admission Control Considerations for Video Interoperability, page 8-22

Guideline and Restrictions for Video Interoperability

The following guidelines and restrictions apply with regard to video interoperability in a Unified CM deployment:

- If H.323 or SCCP protocols are used in conjunction with video interoperability, Unified CM will support only a single H.264 payload and the packetization mode is treated as 0. An example side effect (but not the only one) of this circumstance is the fact that 1080p resolution is not available with these protocols because 1080p requires packetization mode 1.
- If multiple payloads are presented by an H.323 or SCCP endpoint engaged in a video interoperability call, Unified CM will use only the payload with the lowest codec profile. This, in turn, could result in less than the highest supported resolution being selected for the call.
- If a SIP endpoint omits the **level-asymmetry-allowed** parameter in the Session Description Protocol (SDP), Cisco products will assume that the endpoint can support asymmetric resolution transmission. Therefore, different receiving and sending video resolutions could be negotiated during a call.
- If a call is processed with video interoperability while Unified CM is performing protocol interworking with SIP and H.323, the H.323 video endpoint must honor the proposed dynamic payload number specified by the SIP side, which means that no re-negotiation to a different payload would be supported.
- Unified CM will not negotiate Real-Time Transport Control Protocol (RTCP) feedback if the video call invokes a media termination point (MTP) or transcoder.

Quality of Service (QoS) and Call Admission Control Considerations for Video Interoperability

There are no changes to the configuration of regions and locations in Unified CM as a result of video interoperability support. However, regions play a significant role in determining the resolution between groups of endpoints, and they can be used to maximize or minimize the resolution that these devices use when interoperating. The **Max Video Call Bit Rate** field in the regions settings is used to determine the amount of bandwidth and, thus, the resolution that endpoints are able to negotiate.

For further information about QoS and call admission control with native video interoperability, see the section on Call Admission Control Design Recommendations for Video Deployments, page 13-78.

Software-Based Endpoints

A software-based endpoint is an application installed on a client desktop computer that registers and communicates with Cisco call processing platforms for voice and video services. In addition, these endpoint software client applications may provide collaboration features and services such as messaging, presence, directory access, and conferencing. Software-based endpoint desktop client applications include Cisco IP Communicator and Cisco Jabber.

Cisco IP Communicator

Cisco IP Communicator is a Microsoft Windows-based application that provides enterprise IP phone functionality to desktop computers. This application provides enterprise-class IP voice calling for remote users, telecommuters, and other mobile users. Cisco IP Communicator supports both SCCP and SIP signaling protocols for registering and communicating with Cisco call processing platforms. For more information about Cisco IP Communicator, refer to the data sheets and product documentation at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/ip-communicator/index.html

Cisco Jabber Desktop Clients

Cisco Jabber desktop clients enable integration of collaboration services, including audio, video, web collaboration, visual voicemail, and so forth, into a software-based desktop application. Cisco Jabber allows desktop application users to access a variety of communication and collaboration services as provided by back-end collaboration application servers such as Cisco Unified Communications Manager (Unified CM), Cisco IM and Presence, Cisco Unity Connection, Cisco WebEx, and Lightweight Directory Access Protocol (LDAP)-compliant directories. Cisco Jabber is able to leverage IM and presence capabilities provided by either on-premises Cisco IM and Presence or the Cisco WebEx Messenger cloud service.

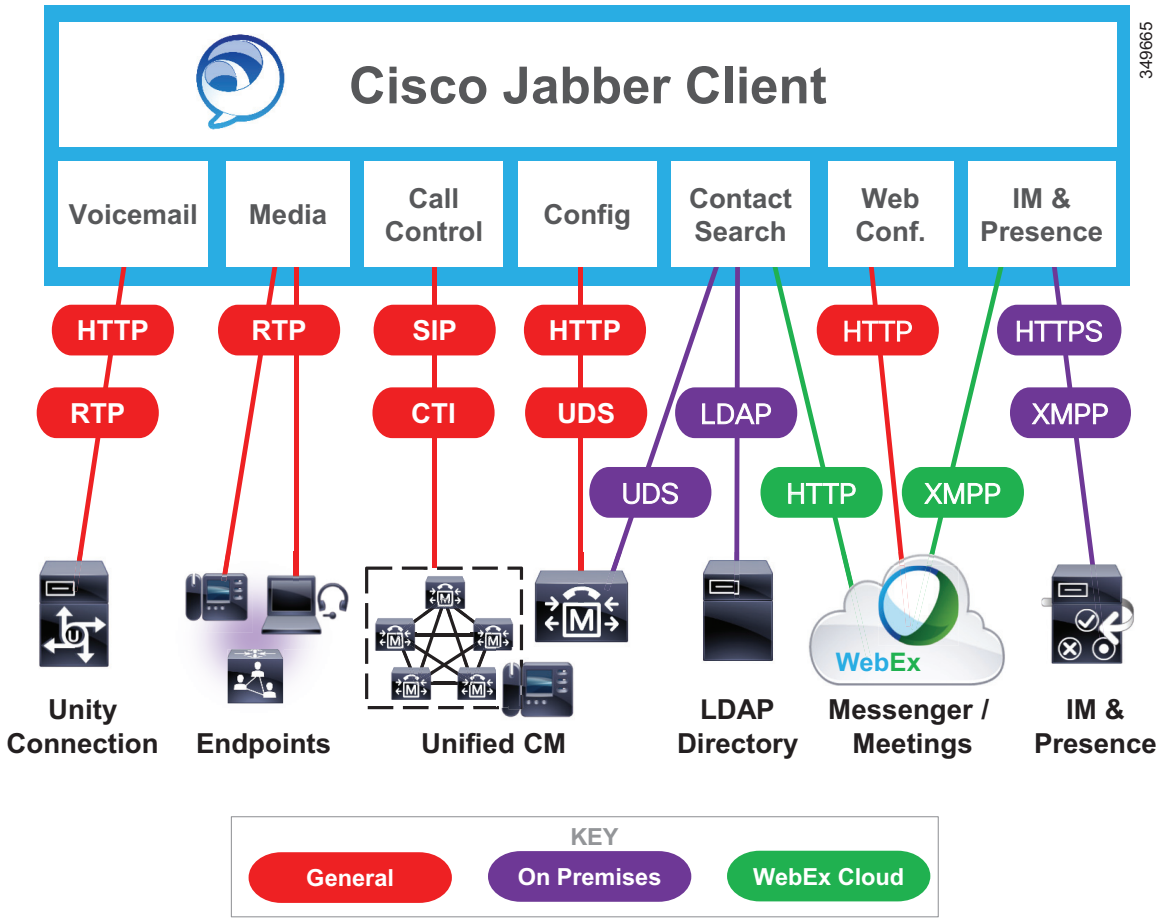
Cisco Jabber Desktop Client Architecture

Cisco Jabber for Windows and Cisco Jabber for Mac use a common set of services to provide various Cisco collaboration features, including instant messaging and presence, audio, video, web collaboration, visual voicemail, and so forth. This common set of services provides a simplified client interface and an abstraction layer that allows access to the following underlying communications services:

- SIP-based call control for voice and video softphone clients from Unified CM
- Deskphone call control and "Click to Call" services from the Unified CM CTI interface
- Voice and video media termination for softphone clients
- Instant messaging and presence services using XMPP, from either the Cisco IM and Presence Service or Cisco WebEx Messenger service. Cisco WebEx Meeting Center also offers hosted collaboration services such as online meetings and events.
- Viewing scheduled audio, video and web conferencing services
- Desktop sharing using either video desktop sharing (BFCP) or WebEx desktop sharing
- Visual voicemail services from Cisco Unity Connection using Internet Message Access Protocol (IMAP) or Representational State Transfer (REST)
- Contact management relying on Unified CM User Data Service (UDS), Microsoft Active Directory, or other supported LDAP directories; or in the case of cloud-based integrations, the WebEx Messenger service
- Microsoft Outlook Integration, which provides user availability status and messaging capabilities directly through the user interface of Microsoft Office applications such as Microsoft Outlook

The ability to communicate and abstract services and APIs, as shown in Figure 8-2, allows the Jabber Desktop Client to coordinate the management of protocols to these services and APIs, handle event notifications, and control the low-level connection logic for local system resources. Depending on the deployment type, some features might not be supported.

Figure 8-2 Cisco Jabber Desktop Client Architecture



Jabber Desktop Clients – Instant Messaging and Presence Services

Instant messaging and presence services for Jabber clients are provided through an XMPP interface. Cisco offers instant messaging and presence services with the following products:

- Cisco IM and Presence
- WebEx Messenger service

The choice between Cisco IM and Presence or WebEx Messenger service for instant messaging and presence services can depend on a number of factors. WebEx Messenger service deployments use a cloud-based service that is accessible from the Internet. On-premises deployments based on Cisco IM and Presence provide the administrator with direct control over their IM and presence platform and also allow presence federation using SIP/SIMPLE to other presence services.

For information on the full set of features supported by each IM and Presence platform, refer to the following documentation:

- Cisco IM and Presence

  https://www.cisco.com/c/en/us/products/unified-communications/unified-presence/index.html

- WebEx Messenger service

  https://www.cisco.com/c/en/us/products/unified-communications/webex-messenger/index.html

Jabber Desktop Clients – Call Control

Cisco Jabber Desktop Clients can operate in one of two modes for call control:

- Softphone Mode — Using audio and video on a computer

  When a Jabber Desktop Client is in softphone mode, it is directly registered to Unified CM as a SIP endpoint for audio and video call control functionality, and it is configured on Unified CM as device type Client Services Framework.

- Deskphone Control Mode — Using a Cisco IP Phone for audio (and video, if supported)

  When a Jabber Desktop Client is in deskphone control mode, it does not register with Unified CM using SIP, but instead it uses CTI/JTAPI to initiate, monitor, and terminate calls, monitor line state, and provide call history, while controlling a Cisco Unified IP Phone. The Cisco CallManager Cisco IP Phone (CCMCIP) or UDS service on Unified CM is used by the Jabber Desktop Client to retrieve a list of devices associated with each user. This list of devices is used by a client in deskphone mode to choose which Cisco IP Phone it wishes to control.

Softphone Mode

When operating in softphone mode, the Jabber Desktop Client is a SIP line-side registered device on Unified CM, utilizing all the call control capabilities and functionality of a Cisco Unified IP Phone, including configuration of registration, redundancy, regions, locations, dial plan management, authentication, encryption, user association, and so forth. The Jabber Desktop Client supports a single line appearance for the user.

The SIP registered device of the Jabber Desktop Client must be factored in as a regular SIP endpoint, like any other SIP registered endpoint, for purposes of sizing calculations for a Unified CM cluster. The Jabber Desktop Client in softphone mode uses the CCMCIP or UDS service to discover its device name for registration with Unified CM.

Deskphone Control Mode

When operating in deskphone control mode, the Jabber Desktop Client uses CTI/JTAPI to provide the ability to place, monitor, and receive calls using Cisco Unified IP Phones. When audio calls are received or placed in this mode, the audio path is through the Cisco Unified IP Phone. For video calls, the video stream can originate and terminate either on the Cisco IP Phone (if it has a camera) or on the computer using a approved camera. The Jabber Desktop Client uses the CCMCIP or UDS service on Unified CM to discover the associated devices of the user.

When using deskphone control mode for the Jabber Desktop Client, factor the CTI scaling numbers into the Unified CM deployment calculations. For additional information about capacity planning, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Jabber Desktop Clients – Audio, Video, and Web Conferencing Services

Access to scheduled conferencing services for clients can be provided through an HTTP interface. Cisco audio, video and web-based scheduled conferencing services can be provided by using the cloud-based WebEx Meeting Center service or a combination of on-premises WebEx Meeting Server for audio and video conferencing services and WebEx cloud-based web conferencing services. For more information about WebEx Meeting Center, refer to the documentation available at:

https://www.cisco.com/c/en/us/products/conferencing/webex-meeting-center/index.html

Jabber Desktop Clients – Contact Management

The Jabber Desktop Client can use one of the following contact sources for contact search and information:

- Cisco Unified CM User database via the User Data Service (UDS)
- LDAP directory integration
- WebEx Messenger service

Contacts can also be stored and retrieved locally using either of the following:

- Jabber Desktop Client Cache
- Local address books and contact lists such as Microsoft Outlook

The Jabber Desktop Client uses reverse number lookup to map an incoming telephone number to a contact, in addition to photo retrieval. The Jabber Desktop Client contact management allows for up to five search bases to be defined for LDAP queries.

Cisco Unified CM User Data Service (UDS)

UDS provides clients with a contact search service on Cisco Unified Communications Manager. You can synchronize contact data into the Cisco Unified CM User database from Microsoft Active Directory or other LDAP directory sources. Clients can then automatically retrieve that contact data directly from Unified CM using the UDS REST interface.

The UDS-to-LDAP Proxy feature is available as an alternate to sourcing contact information from the local Unified CM user database. With UDS-to-LDAP Proxy, contact searches are still handled by UDS but are proxied to the corporate LDAP directory, with UDS relaying results back to the Jabber client. This enables Jabber clients to search a corporate directory that exceeds the maximum number of users supported within the Unified CM database.

LDAP Directory

You can configure a corporate LDAP directory to satisfy a number of different requirements, including the following:

- User provisioning — You can provision users automatically from the LDAP directory into the Cisco Unified Communications Manager database using directory integration. Cisco Unified CM synchronizes with the LDAP directory content so that you avoid having to add, remove, or modify user information manually each time a change occurs in the LDAP directory.
- User authentication — You can authenticate users using the LDAP directory credentials. Cisco IM and Presence synchronizes all the user information from Cisco Unified Communications Manager to provide authentication for client users.
- User lookup — You can enable LDAP directory lookups to allow Cisco clients or third-party XMPP clients to search for contacts in the LDAP directory.

WebEx Directory Integration

Use the WebEx Administration Tool to implement WebEx Directory Integration. WebEx imports a comma-separated value (CSV) file of your enterprise directory information into its WebEx Messenger service. For more information, refer to the latest version of the *Cisco WebEx Messenger Administration Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/webex-messenger/products-installation-guides-list.html

Jabber Desktop Client Cache

The Jabber Desktop Client maintains a local cache of contact information derived from previous directory queries and contacts already listed, as well as the local address book or contact list. If a contact for a call already exists in the cache, the Jabber Desktop Client does not search the directory. If a contact does not exist in the cache, the Jabber Desktop Client performs a directory search.

Directory Search

When a contact cannot be found in the local Jabber Desktop Client cache or contact list, a search for contacts can be made. The WebEx Messenger user can utilize a predictive search whereby the cache, contact list, and local Outlook contact list are queried as the contact name is being entered. If no matches are found, the search continues to query the corporate directory (WebEx Messenger database).

For more information about Cisco Jabber for Windows, refer to the data sheets and product documentation at

https://www.cisco.com/c/en/us/products/unified-communications/jabber-windows/index.html

For more information about the Cisco Jabber for Mac, refer to the data sheets and product documentation at

https://www.cisco.com/c/en/us/products/unified-communications/jabber-mac/index.html

Cisco Spark Desktop Clients

Cisco Spark desktop clients enable persistent cloud-based virtual team spaces that facilitate 1-to-1 and team collaboration. The Cisco Spark desktop client runs on Windows and Mac computers. Cisco Spark allows desktop application users to access collaboration services from the Cisco Collaboration Cloud, including secure and encrypted persistent messaging, voice and video calls over IP, and file sharing, all within virtual one-on-one or group collaboration spaces. The client communicates with the Cisco Collaboration Cloud using HTTPS for messaging and file sharing, while voice and video over IP media uses SRTP.

For proper Cisco Spark client operation, the desktop computer must be able to reach the Internet by connecting to a wired or wireless network (802.11 WLAN or mobile provider data network).

For more information about the Cisco Spark desktop clients, additional feature details, and supported hardware and software versions, refer to the Cisco Spark documentation at

https://support.ciscospark.com/

Cisco UC Integration™ for Microsoft Lync

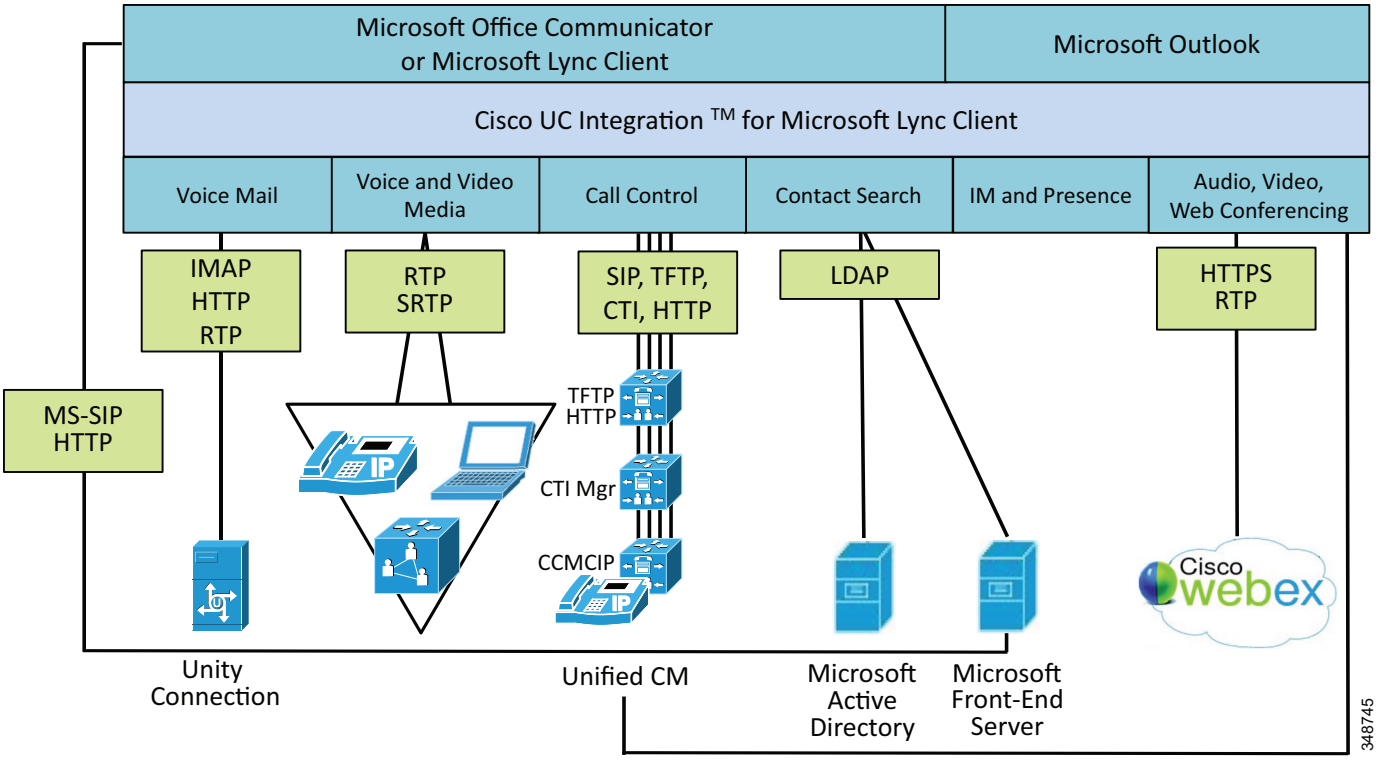
Cisco UC Integration™ for Microsoft Lync clients support a variation of the on-premises deployment models, where IM and presence services are provided by Microsoft Applications instead of Cisco IM and Presence.

Cisco UC Integration™ for Microsoft Lync allows for tightly integrated Cisco Unified Communications services for Microsoft Lync by integrating with underlying Unified Communications services. The solution extends the presence and instant messaging capabilities of Microsoft Lync by providing access to a broad set of Cisco Unified Communications services, including standards-based audio and video, unified messaging, web conferencing, deskphone control, and telephony presence, while delivering a consistent user experience.

Cisco UC Integration™ for Microsoft Lync Architecture

The solution architecture for a Cisco UC Integration™ for Microsoft Lync deployment, shown in Figure 8-3, includes Cisco Unified Communications Manager for audio and video services, Microsoft Office Communications Server 2007 for presence and instant messaging services, Microsoft Active Directory for user account information, Cisco Unified Communications services for PC audio or deskphone control, and Microsoft Lync.

Figure 8-3 Cisco UC Integration™ for Microsoft Lync Architecture



With a deployment of Cisco UC Integration™ for Microsoft Lync, the client utilizes user information from the Office Communications Server Address Book that gets downloaded to the client. The address book is generated and delivered to the clients from the Office Communications Server once the user is enabled for presence and instant messaging. Cisco recommends that administrators populate the user directory number information with an E.164 value (for example, +18005551212) and enable LDAP synchronization and authentication on Unified CM for user account consistency. Cisco UC Integration™ for Microsoft Lync connects to both Cisco Unified CM and Microsoft Active Directory and provides for account credential synchronization rules.

Note With Cisco UC Integration™ for Microsoft Lync, instant messaging and presence services are provided by Microsoft rather than by Cisco Unified Communications services.

Deploying and Configuring Cisco UC Integration™ for Microsoft Lync

When deploying Cisco UC Integration™ for Microsoft Lync, Cisco Unified Communications Manager provides the call control while Microsoft Lync provides the instant messaging and presence.

Cisco UC Integration™ for Microsoft Lync reads its configuration settings from a series of registry entries that the administrator must configure. Cisco recommends pushing these registry configuration settings from Microsoft Active Directory by means of Group Policy to distribute the configuration settings automatically to the client computer. Although Group Policy is the recommended installation mechanism, there are other methods available as well, including third-party software deployment tools, batch files, Vbscript, or manual configuration.

Microsoft Active Directory group policies can be extended using administration templates, and Cisco UC Integration™ for Microsoft Lync provides a template that the administrator can add to provide the group policy support. After the administrative template is loaded, a Cisco UC Integration™ for Microsoft Lync configuration policy can be created by the administrator for the registry configuration settings (TFTP servers, CTI servers, CCMCIP servers, voicemail, and LDAP servers).

The Group Policy Management Console can be used to control how and where these group policies are applied to different organizational units. From a client policy perspective, when you deploy Cisco UC Integration™ for Microsoft Lync, Cisco recommends setting the Microsoft Telephony Mode Policy to **IM and Presence Only** and **DisableAVConferencing**. These client policy changes will allow for only a single set of call options to be displayed in the Microsoft Lync user experience.

A Cisco UC Integration™ for Microsoft Lync deployment also allows for custom presence states to be defined and deployed in the cisco-presence-states-config.xml file that gets installed. However, Cisco recommends that administrators relocate this file to an HTTPs location, such as the Microsoft Office Communications Server, to allow Microsoft Lync to use this custom presence state file based on the following registry location:

HKLM\Software\Policies\Microsoft\Communicator\CustomStateURL

General Deployment Considerations for Software-Based Endpoints

The following sections list important design considerations for deploying software-based endpoints.

Quality of Service

Cisco software-based client applications do mark their traffic at Layer 3 in accordance with QoS marking best practices; however, even when the applications do mark traffic properly, the underlying operating system or hardware might not honor the markings. Given the general unpredictability and unreliability of traffic marking coming from desktop computers, as a general rule these traffic markings should not be trusted. This means that all traffic flows must be re-marked by the network based on protocol and/or port numbers, with real-time traffic flows being marked based on best practices. This includes re-marking of voice-only call media with DSC 46 or PHB EF, video call media (including voice) with DSCP 34 or PHB AF41, and call signaling with DSCP 24 or PHB CS3. These markings along with a properly configured network infrastructure ensure priority treatment for voice-only call media and dedicated bandwidth for video call media and call signaling. In addition to re-marking of software-based endpoint traffic, Cisco recommends using network-based policing and rate limiting to ensure that the software-based endpoint does not consume too much network bandwidth. This can occur when the desktop computer generates too much data traffic or when the endpoint application misbehaves and generates more voice and/or video media and signaling traffic than would be expected for a typical call. In cases where third-party software is used to fully control desktop computer network traffic

marking, administrators may decide to trust desktop computer marking, in which case re-marking of packets would not be required. Network-based policing and rate limiting is still recommended to protect the overall network in case of a misbehaving endpoint.

In the case of software-based Cisco IP Communicator and Cisco Jabber for Windows, appropriate Layer 3 DSCP QoS marking can be applied to audio and video streams based on voice and video media source port numbers using Microsoft Windows group policies.

For more information about Cisco Jabber for Windows QoS with Microsoft Windows group policies, refer to the Quality of Service configuration information in the latest version of *On-Premises Deployment for Cisco Jabber*, available at

https://www.cisco.com/c/en/us/support/unified-communications/jabber-windows/products-installation-guides-list.html

Inter-VLAN Routing

Because software-based endpoints run on a desktop computer usually deployed on a data VLAN, when software-based endpoints are deployed on networks with voice and data VLAN separation, inter-VLAN routing should be configured and allowed so that voice traffic from these endpoints on the data VLAN can reach endpoints on the voice VLAN.

SRST and Enhanced SRST

When deploying Cisco software-based endpoint desktop applications in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. By using SRST or Enhanced SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for software-based endpoints when connectivity to the centralized call processing platform is lost. However, the set of available user-facing features is much smaller when a desktop software-based endpoint is registered to SRST than when the application is registered to Unified CM.

Secure Remote Enterprise Attachment

Cisco software-based endpoints can be securely connected to the enterprise network from remote locations using VPN or VPN-less solutions.

In the case of VPN-based connectivity, software-based endpoints can be located behind a VPN router that creates a secure VPN tunnel to the Cisco Adaptive Security Appliance (ASA) or other VPN head-end concentrator at the enterprise edge. This remote secure connectivity secures not only voice and video media and signaling traffic, but also all traffic coming from the personal computer. As a result, all traffic from the computer traverses the enterprise network edge even if that traffic is ultimately destined for the Internet.

Alternatively, Cisco Jabber desktop clients are able to leverage mobile and remote access functionality of the Cisco Expressway solution. This firewall traversal solution relies on TLS reverse proxy connectivity to the enterprise, as provided by the Cisco Expressway-E and Expressway-C servers for registration to Unified CM call control for voice and video calling and access to enterprise collaboration applications and services such as IM and presence, voicemail, and directory access. For more information about mobile and remote access capabilities of the Cisco Expressway solution, refer to the solution information and product documentation available at

https://www.cisco.com/c/en/us/products/unified-communications/expressway-series/index.html

Dial Plan

Dial plan and number normalization considerations must be taken into account when deploying software-based endpoints. Jabber desktop clients typically use the directory for searching, resolving, and adding contacts. The number that is associated with those contacts must be in a form that the client can recognize, resolve, and dial.

Deployments may vary, depending on the configuration of the directory and Unified CM. In cases where the directory contains E.164 numbering (for example, +18005551212) for business, mobile, and home telephone numbers and Unified CM also contains an E.164 dial plan, the need for additional dial rules is minimized because every lookup, resolution, and dialed event results in an E.164 formatted dial string.

If a Unified CM deployment has implemented a private dial plan (for example, 5551212), then translation of the E.164 number to a private directory number needs to occur on Unified CM. Outbound calls can be translated by Unified CM translation patterns that allow the number being dialed (for example, +18005551212) to be presented to the endpoint as the private number (5551212 in this example). Inbound calls can be translated by means of directory lookup rules. This allows an incoming number of 5551212 to be presented for reverse number lookup caller identification as +18005551212.

Private numbering plan deployments may arise, where the dial plan used for your company and the telephone number information stored in the LDAP directory may require the configuration of translation patterns and directory lookup rules in Cisco Unified Communications Manager to manage number format differences. Directory lookup rules define how to reformat the inbound call ID to be used as a directory lookup key. Translation patterns define how to transform a phone number retrieved from the LDAP directory for outbound dialing.

Translation patterns are used by Unified CM to manipulate the dialed digits before a call is routed, and they are handled strictly by Unified CM. Translation patterns are the recommended method for manipulating dialed numbers. For additional guidelines on translation pattern usage and dial plan management, see the chapter on Dial Plan, page 14-1.

Application dialing rules can be used as an alternative to translation patterns to manipulate numbers that are dialed. Application dialing rules can automatically strip numbers from, or add numbers to, phone numbers that the user dials. Application Dial Rules are configured in Unified CM and are downloaded to the client from Unified CM. Translation patterns are the recommended method for manipulating dialed numbers.

Directory lookup rules transform caller identification numbers into numbers that can be looked up in the directory. A directory lookup rule specifies which numbers to transform based on the initial digits and the length of the number. Directory lookup rules are configured in Unified CM and are downloaded to the client from Unified CM.

Before a call is placed through contact information, the client application removes everything from the phone number to be dialed, except for letters and digits. The application transforms the letters to digits and applies the dialing rules. The letter-to-digit mapping is locale-specific and corresponds to the letters found on a standard telephone keypad for that locale. For example, for a US English locale, 1-800-4UCSRND transforms to 18004827763. Users cannot view or modify the client transformed numbers before the application places the call.

Contact Sources

Cisco Jabber for Windows and Cisco Jabber for Mac default to using Cisco Directory Integration (CDI), which uses service discovery to automatically connect and authenticate to LDAP v3 compatible directories, including Microsoft Active Directory.

For integration with an LDAP directory that requires custom attribute mapping, these attribute mappings can be created in a configuration file that can be downloaded to the client from the Unified CM server.

Cisco Jabber desktop clients also support the Unified CM User Data Service (UDS), which allows a client to search for contacts using the Unified CM user database (which may be synchronized with an LDAP directory). UDS is used automatically by Jabber desktop clients for contact resolution when they are located outside of the corporate firewall and connected via Expressway mobile and remote access.

In addition, Jabber for Windows supports Microsoft Outlook local contact, which allows users to search for contacts that are in the user's Microsoft Outlook client.

Extend and Connect

Cisco Jabber desktop clients support Extend and Connect, which enables users to make and receive calls from Jabber using third-party phones. This allows users to utilize their existing third-party PBX phones while taking advantage of Cisco Collaboration features. There are several modes within Extend and Connect, and each mode requires different trunk usage. The dial plan must be designed carefully for Extend and Connect. For more details on dial plan design, see the chapter on Dial Plan, page 14-1. Extend and Connect is not supported when Jabber clients are outside the corporate network and connected through Expressway mobile and remote access.

OAuth with Refresh Login Flow

Beginning with Cisco Jabber 11.9, client authorization and authentication is facilitated using the OAuth 2.0 authorization framework. This provides for faster login and faster re-authentication during launch and network transitions. Prior to Cisco Unified CM 12.0 and Unified CM 11.5(1) SU3, Cisco Jabber used OAuth only when Single-Sign On (SSO) was enabled within the deployment. The OAuth implementation relies on the Unified CM publisher acting as an authorization server responsible for authenticating and then issuing authorization tokens to clients. This token, along with a refresh token, enables the client to request and gain authorization to collaboration services and to quickly renew an expired authorization token using the refresh token. For additional details on the OAuth 2.0 framework, refer to the section on Authorization Framework, page 16-45.

To leverage OAuth for Jabber client authorization and authentication, the **OAuth with Refresh Login Flow** service parameter must be enabled on Cisco Unified CM, Unified CM IM and Presence, and Unity Connection. Likewise, the **Authorize by OAuth token with refresh** setting must be enabled on Expressway-C for Jabber clients to use OAuth over Expressway Mobile and Remote Access.

For more information on deploying OAuth with Cisco Jabber, refer to the latest version of the white paper on *Deploying OAuth with Cisco Collaboration Solution Release 12.0*, available at

https://www.cisco.com/c/en/us/support/unified-communications/jabber-windows/products-installation-guides-list.html

Wireless Endpoints

Cisco wireless endpoints rely on an 802.11 wireless LAN (WLAN) infrastructure for network connectivity and to provide IP telephony functionality and features. This type of endpoint is ideal for mobile users that move around within a single enterprise location or between enterprise locations or environments where traditional wired phones are undesirable or problematic. Cisco offers the following voice and video over WLAN (VVoWLAN) IP phones:

- Cisco Unified Wireless IP Phones, including the Cisco Unified Wireless IP Phone 8821, 7925G, 7925G-EX, and 7926G
- Cisco IP Phone 8861 and 8865
- Cisco DX Series

All are hardware-based phones with built-in radio antenna. The Cisco Unified Wireless IP Phones 7925G, 7925G-EX, and 7926G enable 802.11b, 802.11g, or 802.11a connectivity to the network. The Cisco Unified Wireless IP Phone 8821 and Cisco IP Phones 8861 and 8865 enable 802.11a, 802.11b, 802.11g, 802.11n, and 802.11ac wireless connectivity while the Cisco DX Series endpoints enable 802.11a, 802.11b, 802.11g, and 802.11n wireless connectivity. The Cisco Unified Wireless IP Phones 7925G, 7925G-EX, and 7926G register and communicate with Cisco call processing platforms using SCCP signaling protocol. The Cisco Unified Wireless IP Phone 8821, Cisco IP Phones 8861 and 8865, and DX Series endpoints use the SIP signaling protocol to register and communicate with Cisco call processing platforms.

For more information about the Cisco Unified Wireless IP Phones, refer to the data sheets and product documentation available at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/unified-ip-phone-7900-series/index.html

For more information about the Cisco IP Phone 8800 Series, refer to the data sheets and product documentation available at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/unified-ip-phone-8800-series/index.html

For more information about the Cisco DX Series endpoints, refer to the data sheets and product documentation available at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/desktop-collaboration-experience-dx600-series/index.html

General Deployment Considerations for Wireless Endpoints

The following sections list important design considerations for deploying wireless endpoints.

Network Radio Frequency Design and Site Survey

Before deploying wireless endpoints, you must ensure your WLAN radio frequency (RF) design minimizes same-channel interference while also providing sufficient radio signal levels and non-adjacent channel overlap so that acceptable voice and video quality can be maintained as the device moves from one location to another. In addition, you must perform a complete WLAN site survey to verify network RF design and to ensure that appropriate data rates and security mechanisms are in place. Your site survey should take into consideration which types of antennas will provide the best coverage, as well as where sources of RF interference might exist. Even when using third-party site survey tools,

Cisco highly recommends that you verify the site survey using the wireless endpoint device itself because each endpoint or client radio can behave differently depending on antenna sensitivity and survey application limitations. Cisco recommends relying on the 5 GHz WLAN band (802.11a/n) whenever possible for connecting wireless endpoints capable of generating voice and video traffic. 5 GHz WLANs provide better throughput and less interference for voice and video calls. Refer to the section on Wireless LAN Infrastructure, page 3-61, for more information about wireless network design.

Security: Authentication and Encryption

When deploying wireless endpoints, it is important to consider the security mechanisms used to control access to the network and to protect the network traffic. Cisco wireless endpoints support a wide range of authentication and encryption protocols including WPA, WPA2, EAP-FAST, PEAP, and so forth. Choose an authentication and encryption method that is supported by the WLAN infrastructure and the endpoint devices you deploy, and one that aligns with IT security policies. In addition, ensure that the authentication and encryption method chosen supports a fast rekeying method such as Cisco Centralized Key Management (CCKM) so that active voice and video calls can be maintained when the device is roaming from one location in the network to another.

Note In dual-band WLANs (those with both 2.4 GHz and 5 GHz bands), it is possible to roam between 802.11b/g and 802.11a with the same SSID, provided the client is capable of supporting both bands. However, with some devices this can cause gaps in the voice or video path. In order to avoid these gaps, use only one band for voice and video communications.

Wireless Call Capacity

When deploying wireless devices and enabling wireless device roaming within the enterprise WLAN, it is also important to consider the device connectivity and call capacity of the WLAN infrastructure. Oversubscription of the WLAN infrastructure in terms of number of devices or number of active calls will result in dropped wireless connections, poor voice and video quality, and delayed or failed call setup. The chances of oversubscribing a deployment of voice and video over WLAN are greatly minimized by deploying sufficient numbers of WLAN access points (APs) to handle required call capacities. AP call capacities are based on the number of simultaneous bidirectional streams that can be supported in a single channel cell area. The general rule for VVoWLAN call capacities is as follows:

- Maximum of 27 simultaneous VoWLAN bidirectional streams per 802.11g/n (2.4 GHz) channel cell with Bluetooth disabled or per 802.11a/n/ac (5 GHz) channel and 24 Mbps or higher data rates enabled.
- Maximum of 8 simultaneous VVoWLAN bidirectional streams per 802.11 g/n (2.4 GHz) channel cell with Bluetooth disabled or per 802.11 a/n/ac (5 GHz) channel cell assuming a video resolution of 720p (high-definition) and video bit rate of up to 1 Mbps.

These call capacity values are highly dependent upon the RF environment, the wireless handset features, and underlying WLAN system features. Actual capacities for a particular deployment could be less.

Note A single call between two wireless endpoints associated to the same AP is considered to be two simultaneous bidirectional streams.

The above capacities are based on voice activity detection (VAD) being disabled and a packetization sample size of 20 milliseconds (ms). VAD is a mechanism for conserving bandwidth by not sending RTP packets while no speech is occurring during the call. However, enabling or disabling VAD, also referred

to as Silence Suppression, is sometimes a global configuration depending on the Cisco call control platforms. Thus, if VAD is enabled for wirelessly attached Cisco Unified IP Phones, then it may be enabled for all devices in the deployment. Cisco recommends leaving VAD (Silence Suppression) disabled to provide better overall voice quality.

At a sampling rate of 20 ms, a voice call will generate 50 packets per second (pps) in either direction. Cisco recommends setting the sample rate to 20 ms for almost all cases. By using a larger sample size (for example, 30 or 40 ms), you can increase the number of simultaneous calls per AP, but a larger end-to-end delay will result. In addition, the percentage of acceptable voice packet loss within a wireless environment decreases dramatically with a larger sample size because more of the conversation is missing when a packet is lost. For more information about voice sampling size, see the section on Bandwidth Provisioning, page 3-52.

Bluetooth Support

The Cisco Unified Wireless IP Phones 8821, 7925G, 7925G-EX, and 7926G, the Cisco IP Phones 8861 and 8865, and the Cisco DX Series endpoints are Bluetooth-enabled devices. The Bluetooth radio or module within these wireless Cisco IP phones enables support for Bluetooth headsets. In addition, as previously mentioned, the Cisco IP Phones 8845, 8851, 8861, 8865, and DX Series endpoints support Intelligent Proximity for Mobile Voice with Bluetooth pairing for hands-free audio and mobile contact and call history sharing. Because Bluetooth devices use the same 2.4 GHz radio band as 802.11b/g devices, it is possible that Bluetooth and 802.11b/g-capable devices can interfere with each other, thus resulting in connectivity issues.

While the Bluetooth and 802.11 WLAN radios coexist natively in the Cisco Unified Wireless IP Phones, Cisco IP Phones 8861, 8865, and Cisco DX Series endpoints, greatly reducing and avoiding radio interference between the Bluetooth and 802.11b/g radio, the Bluetooth radio in these wirelessly attached phones can cause interference for other 802.11b/g and Bluetooth radio devices deployed in close proximity. Due to the potential for interference and disruption of 802.11b/g WLAN voice and video devices (which can result in poor voice and video quality, de-registration, and/or call setup delays), Cisco recommends deploying all WLAN voice and video devices on 802.11a, 802.11n, or 802.11ac, which use the 5 GHz radio band. By deploying wireless phones on the 5 GHz radio band, you can avoid interference caused by Bluetooth devices.

If deploying 802.11 WLAN devices on the 5 GHz radio band is not an option and interference on the 2.4 GHz radio band is causing connectivity, functionality, or voice and video quality issues, consider prohibiting or restricting the use of Bluetooth headsets and Bluetooth dependent features such as Intelligent Proximity for Mobile Voice within these deployments.

For more information on Intelligent Proximity for Mobile Voice and Bluetooth pairing on the Cisco 8851, 8861, and DX Series endpoints, see Intelligent Proximity, page 8-13.

Note Using Bluetooth wireless headsets with the battery-powered Cisco Unified Wireless IP Phones will increase battery power consumption on your phone and will result in reduced battery life.

Note The use of Bluetooth headsets and Bluetooth features such as Intelligent Proximity for Mobile Voice can cause interference and possibly service disruption for adjacent wireless clients and endpoints relying on the 2.4 GHz band (802.11b/g/n).

Quality of Service

When configuring network-level quality of service (QoS), Cisco wireless endpoints (including Cisco Unified Wireless IP Phones, the Cisco IP Phone 8861, and Cisco DX Series endpoints) can be trusted and their packet markings honored. By default these endpoints mark the recommended and appropriate Layer 3 values for voice and video media and call signaling (voice media as DSCP 46 or PHB EF; voice and video media as DSCP 34 or PHB AF41 for a video call, and call signaling as DSCP 24 or PHB CS3). Likewise, these devices mark appropriately at Layer 2 (voice media WMM User Priority (UP) of 6; voice and video media for video call WMM UP 5; call signaling WMM UP 4). With these packet markings, end-to-end voice quality on the converged network will be acceptable.

Despite appropriate packet marking at both Layer 2 and Layer 3, multipurpose devices such as the Cisco DX80 are capable of generating large amounts of non-real-time traffic. As such, concerns are sometimes raised regarding commingling of these devices on the same WLAN SSID or VLAN. While Layer 2 QoS marking and 802.11e WMM work to ensure that more bandwidth and more frequent access to the wireless medium are provided for real-time traffic, in dense or heavily utilized deployments, separating multi-purpose devices such as DX Series endpoints into a separate SSID may provide some relief. However, this separate SSID for multipurpose devices should still be configured with a Platinum QoS profile to ensure that real-time traffic generated by these devices is still given priority treatment across the wireless infrastructure.

SRST and Enhanced SRST

When deploying wireless endpoints in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. By deploying SRST or Enhanced SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for wireless endpoints when connectivity to the centralized call processing platform is lost. However, the set of available user-facing features is much smaller when a wireless endpoint is registered to SRST than when it is registered to Unified CM.

Device Mobility

When wireless endpoints move between locations in a multi-site centralized call processing deployment, the Cisco Unified CM Device Mobility feature may be used to dynamically update the location of the device based on the IP address the device uses to register to Unified CM. This prevents issues with call routing, PSTN egress, and codec and media resource selection typically encountered when devices move between locations. For more information on Device Mobility, see the section on Device Mobility, page 21-14.

For more information about deploying wireless IP endpoints such as the Cisco Unified Wireless IP Phone 7925G, refer to the deployment guides at

https://www.cisco.com/c/en/us/support/collaboration-endpoints/unified-ip-phone-7900-series/products-implementation-design-guides-list.html

For more information about deploying wireless Cisco 8800 Series endpoints, refer to the deployment guide at

https://www.cisco.com/c/en/us/support/collaboration-endpoints/unified-ip-phone-8800-series/products-implementation-design-guides-list.html

For more information about deploying the Cisco DX Series endpoints wirelessly, refer to the deployment guide at

https://www.cisco.com/c/en/us/support/collaboration-endpoints/desktop-collaboration-experience-dx600-series/products-implementation-design-guides-list.html

Mobile Endpoints

Cisco mobile endpoint devices and mobile endpoint client applications register and communicate with Unified CM for voice and video calling services. These devices and clients also enable additional features and services such as enterprise messaging, presence, and corporate directory integration by communicating with other back-end systems such as Cisco Unity Connection, Cisco IM and Presence, and LDAP directories. Cisco offers the following mobile endpoint devices and clients:

- Cisco Jabber for Android and Apple iOS, page 8-37
- Cisco Spark Mobile Clients, page 8-37
- Cisco WebEx Meetings, page 8-38, for Android, BlackBerry and Apple iOS devices
- Cisco AnyConnect Secure Mobility Client, page 8-38, for Android and Apple iOS devices

Cisco Jabber for Android and Apple iOS

The Cisco Jabber mobile clients for Android and Apple iOS devices including the iPhone and iPad enable smartphones and tablets to make and receive enterprise calls using voice and video over IP. The Cisco Jabber mobile client application running on the Android or Apple iOS device registers and communicates with Unified CM using the SIP signaling protocol. The Cisco Jabber mobile client also enables additional features such as corporate directory access, enterprise visual voicemail, XMPP-based enterprise instant messaging and presence, and secure remote attachment with Cisco Expressway mobile and remote access.

For more information about Cisco Jabber for Android, refer to the data sheet and product documentation at

https://www.cisco.com/c/en/us/products/unified-communications/jabber-android/index.html

For more information about Cisco Jabber for iPhone and iPad, refer to the data sheet and product documentation at

https://www.cisco.com/c/en/us/products/unified-communications/jabber-iphone-ipad/index.html

Cisco Spark Mobile Clients

Cisco Spark mobile clients enable persistent cloud-based virtual team spaces that facilitate 1-to-1 and team collaboration. Cisco Spark allows mobile application users to access collaboration services from the Cisco Collaboration Cloud. Cisco Spark for Android, iPad, and iPhone clients provide secure and encrypted persistent messaging, voice and video calls over IP, and file sharing, all within virtual one-on-one or group collaboration spaces. These clients communicate with the Cisco Collaboration Cloud using HTTPS for messaging and file sharing, and SRTP for voice and video over IP media traffic.

For proper Cisco Spark client operation, the mobile client device must be able to reach the Internet by connecting to an 802.11 wireless LAN or mobile provider data network.

For more information about the Cisco Spark mobile clients, additional feature details, and supported hardware and software versions, refer to the Cisco Spark documentation at

https://support.ciscospark.com/

Cisco WebEx Meetings

The Cisco WebEx Meetings mobile client runs on specific Android, Apple iOS, BlackBerry, and Windows Phone mobile smartphones and tablets. This client enables mobile endpoints to participate in Cisco WebEx Meetings with a similar experience as with desktop browser-based Cisco WebEx Meetings. This client enables active participation in Cisco WebEx voice and video conferencing, including the ability to view participant lists and shared content.

For more information about Cisco WebEx mobile clients, refer to the product information at

https://www.cisco.com/c/en/us/products/conferencing/webex-meetings/index.html

Cisco AnyConnect Secure Mobility Client

The Cisco AnyConnect Secure Mobility Client enables secure remote connectivity for Cisco Jabber mobile device clients, enabling persistent enterprise access over mobile data networks and non-enterprise WLANs. This client application provides SSL VPN connectivity for Apple iOS and Android mobile devices through the Cisco AnyConnect VPN solution available with the Cisco Adaptive Security Appliance (ASA) head-end.

For more information on secure remote VPN connectivity using Cisco AnyConnect, refer to the Cisco AnyConnect Secure Mobility Client documentation available at

https://www.cisco.com/c/en/us/support/security/anyconnect-secure-mobility-client/tsd-products-support-series-home.html

Deployment Considerations for Mobile Endpoints and Clients

The following sections list important design considerations for deploying mobile endpoints and clients.

WLAN Design

Because Cisco Jabber mobile clients are often attached to a WLAN, all of the previously mentioned WLAN deployment considerations apply to mobile clients and devices, including WLAN RF design and verification by site survey. In particular, Cisco recommends relying on the 5 GHz WLAN band (802.11a/n/ac) whenever possible for connecting wireless endpoints capable of generating voice and video traffic. 5 GHz WLANs provide better throughput and less interference for voice and video calls. If the 2.4 GHz band is used for mobile clients and devices, Bluetooth should be avoided. Likewise, the WLAN channel cell voice-only and video call capacity numbers covered in the section on Wireless Call Capacity, page 8-34, should be considered when deploying these clients and devices.

Secure Remote Enterprise Attachment

If appropriately deployed, Cisco mobile endpoints and clients can also connect to the enterprise from remote locations by using public or private 802.11 Wi-Fi hot spots or over the mobile data network. In these scenarios, mobile endpoints and clients can be securely connected using VPN or VPN-less solutions. In the case of VPN, the Cisco AnyConnect mobile VPN client can be used to connect the device or client to the enterprise with a secure SSL tunnel.

One important consideration for Cisco Jabber and Cisco AnyConnect deployments is the traffic being secured. When using the Cisco AnyConnect mobile VPN client on a mobile device with Cisco Jabber, the default behavior is that all traffic to and from the device is sent via the encrypted VPN tunnel and

into or through the enterprise. This might not be desirable in all deployments. In the case of Cisco Jabber, the preferred behavior may be to send only the Jabber-specific traffic through the enterprise via the VPN tunnel, while all other traffic is sent outside the tunnel. This can be accomplished by using the split-tunnel feature, which enables administrators to specify which traffic (based on destination subnets) traverses the VPN tunnel and which traffic goes in the clear. To secure just the Jabber traffic, administrators must configure for inclusion in the tunnel the IP subnets of the Cisco Unified Communications Manager cluster, IM and Presence cluster, voicemail server, directory server, and Trivial File Transfer Protocol (TFTP) server as well as the IP subnets for any endpoints they might connect with. Hence, the split-include policy should include the corporate network IP address range. Sometimes the IP space of a large company is not contiguous because of acquisitions and other events, so this configuration might not be applicable for all deployments.

For more information on Cisco Jabber and Cisco AnyConnect with split-tunnel includes, refer to the *Cisco AnyConnect Deployment Guide for Cisco Jabber*, available at

https://www.cisco.com/c/dam/en/us/products/collateral/security/asa-5500-series-next-generation-firewalls/guide_c07-717020.pdf

For VPN-less connectivity, Cisco Jabber mobile clients are able to leverage mobile and remote access functionality of the Cisco Expressway solution. This firewall traversal solution relies on TLS reverse proxy connectivity to the enterprise, as provided by the Cisco Expressway-E and Expressway-C servers for registration to Unified CM call control for voice and video calling and access to enterprise collaboration applications and services such as IM and presence, voicemail, and directory access. For additional information about mobile and remote access capabilities of the Cisco Expressway solution, refer to the solution information and product documentation available at

https://www.cisco.com/c/en/us/products/unified-communications/expressway-series/index.html

Quality of Service

Cisco mobile client applications and devices generally mark Layer 3 QoS packet values in accordance with Cisco collaboration QoS marking recommendations. This includes marking voice-only call media traffic with DSCP 46 or PHB EF, video call media (including voice) traffic with DSCP 34 or PHB AF41, and call signaling traffic with DSCP 24 or PHB CS3. Despite appropriate mobile client and device application Layer 3 packet marking, Layer 2 802.11 WLAN packet marking (User Priority, or UP) presents further challenges. Some devices may appropriately mark wireless Layer 2 802.11 User Priority (UP) values (voice-only call media UP 6, video call media UP 5, and call signaling UP 3). However, because Cisco mobile clients run on a variety of mobile devices, Layer 2 wireless QoS marking is inconsistent and therefore cannot be relied upon to provide appropriate treatment to traffic on the WLAN. In deployments with Cisco Unified Wireless LAN Controllers, enabling wireless SIP call admission control (CAC) might provide some relief for incorrect or nonexistent Layer 2 WLAN marking. SIP CAC utilizes media session snooping and ensures that downstream voice and video frames are prioritized and/or treated correctly. Even assuming appropriate mobile client application Layer 3 or even Layer 2 packet marking, mobile devices present many of the same challenges as desktop computers in terms of generating many different types of traffic, including both data and real-time traffic. Given this, mobile devices generally fall into the untrusted category of collaboration endpoints. For deployments where mobile client devices are not considered trusted endpoints, packet re-marking based on traffic type and port numbers is required to ensure that network priority queuing and dedicated bandwidth are applied to appropriate traffic. In addition to re-marking the mobile device traffic, Cisco recommends using network-based policing and rate limiting to ensure that the mobile client devices do not consume too much network bandwidth.

Note Mobile clients and devices may attach remotely to the enterprise using Cisco AnyConnect client over the mobile data network or public or private Wi-Fi hot spots. Because these connections traverse the Internet, there is no end-to-end QoS on the IP path and therefore all traffic is treated as best-effort. Voice and video quality cannot be guaranteed over these types of connections.

SRST and Enhanced SRST

When deploying mobile endpoints and clients such as Cisco Jabber for iPhone in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. Cisco Jabber mobile clients do not support SRST; however, because most Cisco Jabber mobile clients run on smartphones with cellular voice radios, users may still be able to make call using the mobile provider network.

For additional design and deployment information about Cisco Jabber mobile clients, refer to the section on Cisco Mobile Clients and Devices, page 21-76.

Intelligent Proximity

As previously mentioned, Intelligent Proximity refers to features that leverage proximity-based connections between Cisco hardware endpoints and mobile devices.

Intelligent Proximity for Mobile Voice capabilities available on the Cisco 8851, 8861, and DX Series endpoints rely on Bluetooth pairing between the IP endpoint and a cellular or smart phone, enabling hands-free audio and mobile contact and call history sharing.

As indicated previously, Intelligent Proximity for Mobile Voice on the 8851, 8861, and DX Series endpoints and the Unified Mobility feature set are compatible. Further, Intelligent Proximity for Mobile Voice on the 8851 and 8861 IP Phones is also compatible with Cisco Jabber, enabling audio-playout on the IP Phone 8851 and 8861 while the video is played on the Jabber client device.

For more information on Intelligent Proximity and Bluetooth pairing, see Intelligent Proximity, page 8-13.

Contact Sources

Cisco Jabber for Android and iOS defaults to using Cisco Directory Integration (CDI), which relies on integration to LDAP v3 compatible directories, including Microsoft Active Directory.

For integration with an LDAP directory that requires custom attribute mapping, these attribute mappings can be created in a configuration file that can be downloaded to the client from the Unified CM server.

Cisco Jabber mobile clients also support the Unified CM User Data Service (UDS), which allows a client to search for contacts using the Unified CM user database (which may be synchronized with an LDAP directory). UDS is used automatically by Jabber mobile clients for contact resolution when they are located outside of the corporate firewall and connected via Expressway mobile and remote access.

The UDS-to-LDAP Proxy feature is available as an alternate to sourcing contact information from the local Unified CM user database. With UDS-to-LDAP Proxy, contact searches are still handled by UDS but are proxied to the corporate LDAP directory, with UDS relaying results back to the Jabber client. This enables Jabber clients to search a corporate directory that exceeds the maximum number of users supported within the Unified CM database.

OAuth with Refresh Login Flow

Beginning with Cisco Jabber 11.9, client authorization and authentication is facilitated using the OAuth 2.0 authorization framework. This provides for faster login and faster re-authentication during launch and network transitions. Prior to Cisco Unified CM 12.0 and Unified CM 11.5(1) SU3, Cisco Jabber used OAuth only when Single-Sign On (SSO) was enabled within the deployment. The OAuth implementation relies on the Unified CM publisher acting as an authorization server responsible for authenticating and then issuing authorization tokens to clients. This token, along with a refresh token, enables the client to request and gain authorization to collaboration services and to quickly renew an expired authorization token using the refresh token. For additional details on the OAuth 2.0 framework, refer to the section on Authorization Framework, page 16-45.

To leverage OAuth for Jabber client authorization and authentication, the **OAuth with Refresh Login Flow** service parameter must be enabled on Cisco Unified CM, Unified CM IM and Presence, and Unity Connection. Likewise, the **Authorize by OAuth token with refresh** setting must be enabled on Expressway-C for Jabber clients to use OAuth over Expressway Mobile and Remote Access.

For more information on deploying OAuth with Cisco Jabber, refer to the latest version of the white paper on *Deploying OAuth with Cisco Collaboration Solution Release 12.0*, available at

https://www.cisco.com/c/en/us/support/unified-communications/jabber-windows/products-installation-guides-list.html

Apple Push Notification Service (APNs)

Cisco Jabber for iPhone and iPad 11.9 and later provides support for Apple Push Notification service (APNs) for receiving incoming call and message notifications when the client is running in the background.

Previously, like other Jabber clients, the Jabber for Apple iOS client (iPhone and iPad) leveraged periodic direct IP socket keepalives to maintain connectivity for voice and video over IP (VVoIP) and IM and presence services when the client moved to the background. Because Apple is deprecating the direct IP socket method for notifications, APNs will soon be required for sending notifications to Jabber for iOS clients running in the background on the Apple iOS device.

Cisco Spark for Apple iOS also supports APNs for receiving incoming call and message notifications when the client is running in the background.

Note APNs has *no* impact on non-iOS Cisco Spark clients (Cisco Spark for Android, Cisco Spark for Windows, and Cisco Spark for Mac) and non-iOS Cisco Jabber clients (Jabber for Android, Jabber for Windows, and Jabber for Mac).

Refer to the section on Apple Push Notification Service (APNs) for Cisco Jabber for iPhone and iPad, page 21-99, for more information about APNs for Cisco Jabber clients.

Cisco Virtualization Experience Media Engine

The Cisco Virtualization Experience Media Engine (VXME) provides an integral collaboration software component by extending the Cisco Jabber collaboration experience to a Virtual Desktop Infrastructure (VDI) environment. VXME is a software package installed on a local platform (a thin client), and it allows users to enhance their VDI sessions to include locally terminated voice and video real-time communications, bypassing real-time media routing through the virtual desktop while allowing for a fully integrated user experience. The hosted virtual desktop is supported with Citrix XenDesktop, Citrix XenApp Published Desktop, or VMware View, through locally installed Citrix Receiver or VMware View Client, respectively. Regardless of the host VDI platform, a user has a consistent voice, video, and virtual desktop experience using Cisco Jabber on the virtual desktop with fully integrated accessories enabled for Unified Communications and seamless integration with VXME.

For more information on Cisco Virtualization Experience Media Engine (VXME), refer to the data sheet and product documentation at

https://www.cisco.com/c/en/us/products/collaboration-endpoints/virtualization-experience-media-engine/index.html

Deployment Considerations for Cisco Virtualization Experience Media Engine

The following sections list important design considerations for deploying Cisco Virtualization Experience Media Engine (VXME).

Quality of Service

No additional configuration is required for Cisco Virtualization Experience Media Engine (VXME) if the network is set up for 8021.q Dual VLAN. If the network is not setup for 802.1q Dual VLAN, QoS will be best-effort and the thin client should be placed in the data VLAN. For details on traffic marking, refer to the QoS design guides available at

https://www.cisco.com/c/en/us/solutions/enterprise/design-zone-ipv6/design-guide-listing.html

Call admission control for voice and video follow existing Cisco Unified IP Phone guidelines, and bandwidth controls for the virtual desktop are provided through the connection broker settings.

SRST and Enhanced SRST

When deploying Cisco VXME in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. If Jabber is running in deskphone control mode, during a WAN failure the hosted virtual desktop (HVD) where the Cisco Jabber client runs will continue to have contact with the Cisco Unified CM co-located in the data center. However, Cisco Unified CM connectivity to the desktop phone paired with the VXC zero client will be lost. By using Survivable Remote Site Telephony (SRST) or Enhanced SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for the desktop phones paired with the VXC clients to the centralized call processing platform is lost.

VXME does not support SRST or Enhanced SRST.

Third-Party IP Phones

Some third-party IP phones and devices may be integrated with Cisco call control to provide basic IP telephony functionality, as described in this section.

Third-Party SIP IP Phones

Third-party phones have specific local features that are independent of the call control signaling protocol, such as features access buttons (fixed or variable). Basic SIP RFC support allows for certain desktop features to be the same as on Cisco Unified IP Phones and also allows for interoperability of certain features. However, these third-party SIP phones do not provide the full feature functionality of Cisco Unified IP Phones.

Cisco works with key third-party vendors who are part of the Cisco Developer Network and who are developing solutions that leverage Cisco Unified CM and Unified CME SIP capabilities. For example, Tenacity Operating provides a software-based endpoint called accessaphone ipTTY, which enables terminal teletype (TTY) or text-based communications for IP telephony. This software-based endpoint can register and communicate with Cisco Unified CM as a third-party SIP phone.

For more information on Cisco's line-side SIP interoperability, refer to the Cisco Unified Communications Manager programming guides at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-programming-reference-guides-list.html

For more information on the Cisco Developer Network and third-party development partners, refer to the information available on the Cisco Developer Community at

https://developer.cisco.com

High Availability for Collaboration Endpoints

To stay in service even during failure of the call control platform, Cisco endpoints are capable of being configured with multiple server nodes or servers for registration and call control service redundancy.

In the case of Cisco Unified CM call control, either through direct configuration or through DHCP during the boot-up phase, collaboration endpoints can accept and process more than one TFTP server address. In case the primary TFTP server is down when the endpoint boots up, the endpoint can get its configuration files from the secondary TFTP server.

Each of the endpoints is also associated with a device pool. The device pool contains a Unified CM Group that has one or more Unified CM subscribers. A list of these subscribers is sent to the endpoints in their configuration files. The endpoints attempt to register with the first (the primary) subscriber in the list. If that Unified CM subscriber is unavailable, the endpoint attempts to register with the second subscriber in the list (the secondary), and so on. Once registered to a subscriber, an endpoint can fail-over to another subscriber in the priority list in the Unified CM Group if the current subscriber fails. When a higher-priority subscriber comes back up, the endpoint will re-register to it.

In the case of those endpoints leveraging CTI, for example Cisco Jabber desktop client running in deskphone control mode, CTI service redundancy is required. In those cases, multiple Unified CM nodes should be running the Cisco CTIManager service so that Unified CM Group configuration will provide client failover from the primary to a secondary CTIManager node in the event of a node failure.

To protect against network failure for endpoints located across a WAN from the Unified CM cluster, a locally available Cisco Integrated Services Router (ISR) or other Cisco IOS router with SRST or Enhanced SRST may also be configured in the list of servers with which the endpoint may register. In case of a WAN failure, the endpoints register to the SRST router and provide uninterrupted voice

telephony services (although the set of features they support in SRST mode might be smaller). Note that some endpoints, including Cisco Jabber and Cisco TelePresence System video endpoints, do not support SRST.

Capacity Planning for Collaboration Endpoints

Cisco call control platforms support the following high-level endpoint capacities:

- A Cisco Unified CM cluster, even when deployed as part of Cisco Business Edition 7000, supports a maximum of 40,000 SCCP or SIP endpoints.
- When deployed as part of Cisco Business Edition 6000, a Cisco Unified CM cluster supports a maximum of 2,500 SCCP or SIP endpoints, depending on the server type.
- Cisco Business Edition 4000 supports a maximum of 200 SIP endpoints.
- Cisco Unified CM Express supports a maximum of 450 SCCP or SIP endpoints.
- Cisco Expressway-C and Expressway-E cluster pairs support a maximum of 10,000 remote endpoint proxy registrations.

The above numbers are nominal maximum capacities. The maximum number of endpoints that the call control platform will actually support depends on all of the other functions that the platform is performing, the busy hour call attempts (BHCA) of the users, and so forth, and the actual capacity could be less than the nominal maximum capacity. Unified CM CTI capacity must also be considered when sizing the system to ensure that Jabber desktop clients and other deskphone control applications have sufficient CTI capacity for operation. For more information on CTI sizing, refer to the chapter on Collaboration Solution Sizing Guidance, page 25-1.

In addition to call control platform capacity, network capacity must be considered with regard to bandwidth and call capacity. Of particular concern are 802.11 wireless attached devices such as the Cisco Unified Wireless IP Phone 7925G or an Android smartphone running Cisco Jabber, where network endpoint capacity is not determined by the number of physical ports but by the amount of bandwidth and throughput available on the shared wireless network. See Wireless Call Capacity, page 8-34, for voice and video call capacities per 802.11 channel cell.

For more information on endpoint capacity with Cisco call control, including platform-specific endpoint capacities per node, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Design Considerations for Collaboration Endpoints

The following list summarizes high-level design recommendations for deploying Cisco endpoints:

- Analog gateways are available both as standalone devices and as integrated interface modules on Cisco IOS multiservice routers, and both types can be used within the same deployment. Select the analog gateway or gateways that meet analog port density requirements across company locations. Ensure that appropriate port capacity is provided for all locations in order to accommodate the required analog devices.
- Enable the role of **Standard CTI Allow Control of Phones supporting Connected Xfer and conf** for the end-user configuration associated with the device in order to enable CTI monitoring and control of Cisco IP Phone 8800 Series and Cisco DX Series endpoints. Only after this role has been enabled can CTI applications monitor or control these phones.

- To minimize endpoint firmware upgrade times over the WAN to remote branches, consider deploying a local TFTP server at the remote location and point endpoints located in that branch to this local TFTP server using the **load server** parameter. Alternatively, consider the use of the Peer File Sharing (PFS) feature when all or most of the devices at a particular remote location are the same phone model.
- Cisco Unified IP desk phones can be powered by power over Ethernet (PoE) when plugged into inline power-capable switches or when deployed with an inline power injector. Consider the use of inline power to reduce downtime and eliminate the need for an external power supply and wall power outlet.
- When deploying Cisco endpoints in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. By using SRST or Enhanced SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for the desk phones when connectivity to the centralized call processing platform is lost. However, the set of available user-facing features is much smaller when a device is registered to SRST than when the phone is registered to Unified CM.
- For deployments with network voice and data VLAN separation, ensure that inter-VLAN routing has been configured and allowed so that Cisco software-based endpoints that run on desktop computers usually connected to data VLANs can communicate with endpoints on the voice VLAN. This is also important for endpoints on the voice VLAN that may be dependent on data VLAN-based resources that provide services such as directory and management.
- A WLAN site survey must be conducted to ensure appropriate RF design and to identify and eliminate sources of interference prior to deploying wireless and mobile endpoints capable of generating real-time traffic on the wireless network. This is necessary to ensure acceptable voice and video quality for calls traversing the WLAN.
- Select a WLAN authentication and encryption method that not only adheres to company security policies but also enables fast rekeying or authentication so that audio and video calls are not interrupted when wireless endpoints move from one location to another.
- Cisco recommends relying on the 5 GHz WLAN band (802.11a/n/ac) whenever possible for connecting wireless endpoints and mobile client devices capable of generating voice and/or video traffic. 5 GHz WLANs provide better throughput and less interference for voice and video calls. If the 2.4 GHz band is used for connecting wireless client devices and endpoints, Bluetooth should be avoided.
- Provide appropriate network and call control capacity to support the number of endpoints deployed. First, consider the endpoint registration and configuration capacities per call control platform: maximum of 40,000 endpoints per Unified CM cluster, even when deployed as part of Cisco Business Edition 7000; 2,500 endpoints per cluster when deployed as part of Cisco Business Edition 6000; 200 SIP endpoints when deployed on Cisco Business Edition 4000; or 10,000 remote endpoint registrations over Cisco Expressway. Next, consider call capacities per wireless channel cell for wireless attached endpoints, and the maximum of 27 bidirectional voice-only streams or maximum of 8 simultaneous voice and video streams or calls per WLAN channel cell.
- Ensure that the end-to-end network infrastructure has been configured with appropriate QoS policies, including marking and re-marking as appropriate, trust boundaries, queuing with both priority and dedicated bandwidth queues, rate limiting, and policing, so that collaboration endpoints deliver high-quality voice and video to end users.

Call Processing

Revised: March 1, 2018

The handling and processing of voice and video calls is a critical function provided by IP telephony systems. This functionality is handled by some type of call processing entity or agent. Given the critical nature of call processing operations, it is important to design unified communications deployments to ensure that call processing systems are scalable enough to handle the required number of users and devices and are resilient enough to handle various network and application outages or failures.

This chapter provides guidance for designing scalable and resilient call processing systems with Cisco call processing products. These products include Cisco Unified Communications Manager (Unified CM) and Cisco Unified Communications Manager Express (Unified CME). The discussions focus predominately on the following factors:

- Scale — The number of users, locations, gateways, applications, and so forth
- Performance — The call rate
- Resilience — The amount of redundancy

Specifically, this chapter focuses on the following topics:

- Call Processing Architecture, page 9-2

  This section discusses general call processing architecture and the various call processing hardware options. This section also provides information on Unified CM clustering.

- High Availability for Call Processing, page 9-13

  This section examines high availability considerations for call processing, including network redundancy, server redundancy, and load-balancing.

- Capacity Planning for Call Processing, page 9-23

  This section provides an overview of sizing for call processing deployments.

- Design Considerations for Call Processing, page 9-26

  This section provides a summarized list of high-level design guidelines and best practices for deploying call processing.

- Computer Telephony Integration (CTI), page 9-28

  This section explains the Cisco Computer Telephony Integration (CTI) architecture and discusses CTI components and interfaces, CTI functionality, and CTI provisioning and capacity planning.

- Integration of Multiple Call Processing Agents, page 9-36

  This section discusses the integration of multiple call processing agents, which is typically done with Cisco Unified CM Session Management Edition (SME). It also covers direct integration of Cisco Unified CM with Cisco Unified Communications Manager Express (Unified CME).

What's New in This Chapter

Table 9-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 9-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Minor updates and corrections	Various sections of this chapter	March 1, 2018

Call Processing Architecture

In order to design and deploy a successful Unified Communications system, it is critical to understand the underlying call processing architecture that provides call routing functionality. This functionality is provided by the following Cisco call processing agents:

- Cisco Unified Communications Manager (Unified CM)

  Cisco Unified CM provides call processing services for small to very large single-site deployments, multi-site centralized call processing deployments, and/or multi-site distributed call processing deployments. Unified CM is at the core of a Cisco Collaboration solution, and it serves as a foundation to deliver voice, video, TelePresence, IM and presence, messaging, mobility, web conferencing, and security.

  Access to the enterprise collaboration network and to Unified CM from the internet to enable remote access and business-to-business secure telepresence and video communications, is also available through different collaboration edge solutions such as VPN and Cisco Expressway.

- Cisco TelePresence Video Communication Server (VCS)

  Cisco TelePresence VCS is a video application that provides video endpoint registration, call processing, and bandwidth management for SIP and H.323 endpoints. VCS acts as a SIP registrar, a SIP proxy server, an H.323 gatekeeper, and a SIP-to-H.323 gateway server to provide interworking between SIP and H.323 devices. Cisco TelePresence VCS also provides external communications using NAT/firewall traversal when combined with the VCS Expressway.

  Cisco recommends deploying Unified CM as the main call processing agent for all endpoints, including TelePresence endpoints and room-based TelePresence conferencing systems that support SIP, and use VCS only for full-featured interoperability with H.323 telepresence endpoints or integration with third-party video endpoints. This is to avoid the dial plan and call admission control complexities that dual call control introduces. Therefore, this chapter does not provide many details on VCS. For more information on VCS, refer to the Cisco Collaboration System 10.*x* SRND or the Cisco VCS product documentation.

- Cisco Business Edition 4000 and Cisco Unified Communications Manager Express (Unified CME)

  Cisco Business Edition 4000 (BE4K) is a new on-premises, completely cloud-managed audio telephony platform optimized for small to medium businesses. Powered by Cisco Unified Communications Manager Express (Unified CME) and Cisco Unity Express Virtual (vCUE), Business Edition 4000 provides affordable integrated IP telephony and voicemail solutions for up to 200 devices. As with Cisco Business Edition 6000 and 7000, Business Edition 4000 simplifies the quoting and ordering process and allows for rapid deployments by providing pre-configured hardware, pre-installed licenses, and pre-loaded Cisco Collaboration applications.

  Cisco Business Edition 4000 and Cisco Unified CME provide call processing services for small single-site deployments and larger distributed multi-site deployments. Cisco Unified CME also provides call processing services for deployments in which a local call processing entity at a remote site is needed to provide backup capabilities for a centralized call processing deployment of Cisco Unified CM.

Cisco Unified Communications Manager (Unified CM) and Cisco TelePresence Video Communication Server (VCS) are available as standard Cisco Collaboration products or through Cisco Business Edition 6000 and Cisco Business Edition 7000, which are packaged collaboration solutions that include call processing services and other services such as messaging, conferencing, and contact center.

The Cisco Business Edition 6000 and 7000 solutions simplify the quoting/ordering process and accelerate deployments by providing pre-configured hardware, pre-installed licensed hypervisor, and pre-loaded and/or pre-installed Cisco Collaboration applications. Cisco Business Edition 6000M and Cisco Business Edition 6000H are targeted for deployments with up to 1,000 users. Cisco Business Edition 7000 is targeted for deployments with more than 1,000 users. The design and sizing of the Cisco Collaboration applications have been simplified with Cisco Business Edition 6000. With Cisco Business Edition 7000, however, normal Unified CM design and sizing guidelines apply.

Call Processing Virtualization

Virtualization enables multiple Cisco Collaboration "servers" or "virtual machines" to run on one physical server. The Cisco Collaboration servers or virtual machines are also referred as VMs, nodes, or instances in this document.

This architecture has obvious benefits over traditional deployments where the applications are running directly on the hardware platform. For example, costs (such as server, electricity, cooling, and rack space costs) can be reduced significantly, and the operation and maintenance of the hardware platforms can be simplified. Virtualization is enabled by a hypervisor that is installed directly on the physical server and that manages the virtual machines. The hypervisor that is required with Cisco Collaboration is the VMware ESXi Hypervisor.

Each virtual machine has associated virtual hardware resources such as virtual CPU, virtual memory, and virtual disk. Those resources are defined for each Collaboration application in predefined templates that are distributed through an Open Virtualization Archive (OVA), an open standards-based method for packaging and distributing virtual machine templates. For many of the Cisco Collaboration applications, in order to provide different capacity options, several VM configuration options are available when deploying an OVA. OVAs must be used when installing a Cisco Collaboration application, not only to define the correct virtual hardware resources but also to ensure that the virtual disks are not misaligned with the host physical disks, which would impact the storage performance.

The virtualization support for the Cisco Collaboration call processing agents is as follows:

- Cisco Unified CM runs only as a virtual application; it cannot be deployed directly on a Cisco UCS server, for example.
- Cisco Unified CME runs within the Cisco IOS or IOS-XE software on Cisco Integrated Services Routers and does not support virtualization.
- Cisco Business Edition 4000 runs within the Cisco IOS-XE software on a Cisco 4321 Integrated Services Router (ISR) and does not support virtualization.

For more information on the considerations for designing and deploying virtualization of Cisco Unified Communications applications, refer to the information available at

https://www.cisco.com/go/virtualized-collaboration

Call Processing Hardware

There are three types of hardware options for Cisco Unified CM: Tested Reference Configurations, Cisco Business Edition 6000 and 7000, and Specifications-based hardware.

- Tested Reference Configuration (TRC)

  TRCs are selected hardware configurations based on the Cisco Unified Computing System (UCS) servers. They have a fixed hardware configuration, and they are tested and validated with Cisco Collaboration applications for specific advertised performance, capacity, and application co-residency scenarios. They are intended for customers who require explicitly validated infrastructure and/or customers who are not necessarily experienced with virtualization.

  The hardware configuration for each TRC is well defined, and allowed deviation from this hardware configuration is very limited. For example, changing the CPU model or number of cores, or changing the RAID configuration of a TRC, would change the server qualification, and the server would not be considered as a TRC anymore but rather as specifications-based hardware.

- Cisco Business Edition 6000 and 7000

  Cisco Business Edition 6000 and 7000 are packaged collaboration solutions that include the hardware platform, virtualization software, and Cisco applications. The hardware platform is pre-configured (for instance, firmware, drivers and the RAID controller are pre-configured at the factory). Just like the TRC, the hardware platform is tested and validated with Cisco Collaboration applications for specific capacity and performance.

  Cisco Business Edition 6000 is available with two hardware platform options: BE6000M and BE6000H. Cisco Business Edition 7000 also is available with two hardware platform options: BE7000M and BE7000H.

  For more details on the TRC and Cisco Business Edition 6000 and 7000 hardware platforms, refer to the documentation at https://www.cisco.com/go/virtualized-collaboration.

- Specifications-based hardware

  Specifications-based hardware (sometimes simply referred as "specs-based") provides more flexible hardware configurations. For example, it allows you to select a platform based on a Cisco UCS TRC and to change the CPU model, number of cores, and RAID configuration, and/or to use an iSCSI or NAS storage. If desired, it also allows you to use a server vendor other than Cisco. Any specifications-based hardware server, whether it is Cisco or not, must be listed in the following *VMware Compatibility Guide*:

  https://www.vmware.com/resources/compatibility/search.php

While specification-based hardware provides more flexible hardware configurations, some requirements must still be met. For example, there are requirements around the CPU model and minimum CPU speed, and vCenter is required in order to collect logs and statistics. With specifications-based hardware, it is important to understand that the hardware configuration has not been explicitly validated by Cisco with Cisco Collaboration applications. Therefore Collaboration applications cannot provide prescriptive guidance on hardware compatibility and cannot be guaranteed, and performance of the Cisco Collaboration applications is for guidance only (refer to the Troubleshooting TechNote at https://www.cisco.com/c/en/us/support/docs/voice-unified-communications/unified-communications-system/115955-uc-specs-tshoot-00.html).

To obtain guidance on the performance of Cisco Collaboration applications with specifications-based hardware, use the TRCs or Cisco Business Edition 6000 and 7000 hardware platforms as references. For more information, refer to the documentation at https://www.cisco.com/go/virtualized-collaboration.

Cisco Unified CME runs on Cisco Integrated Services Routers (ISR) such as the Cisco 2900, 3900, or 4000 Series ISRs. Cisco Unified CME does not run as a virtual application or as part of Cloud Services Router 1000V. Cisco Business Edition 4000 runs on Cisco 4321 Integrated Services Router (ISR) only. The voicemail powered by Cisco Unity Express Virtual (vCUE) can be run only within the Service Container of the Cisco ISR 4321 and not on a UCS-E module.

Determining the appropriate call processing type and platform for a particular deployment will depend on the scale, performance, and redundancy required. In general, Unified CM provides a very wide range of capacity options and higher availability, while Cisco Unified CME provides lower levels of capacity and redundancy. For specifics regarding redundancy and scalability, see the sections on High Availability for Call Processing, page 9-13, and Capacity Planning for Call Processing, page 9-23.

Unified CM Cluster Services

While Cisco Unified CME is a standalone call processing application, Unified CM supports the concept of clustering. The Unified CM architecture enables a group of server nodes to work together as a single call processing entity or IP PBX system. This grouping of server nodes is known as a *cluster*. A cluster of Unified CM server nodes may be distributed across an IP network, within design limitations, allowing for spatial redundancy and, hence, resilience to be designed into the Unified Communications System.

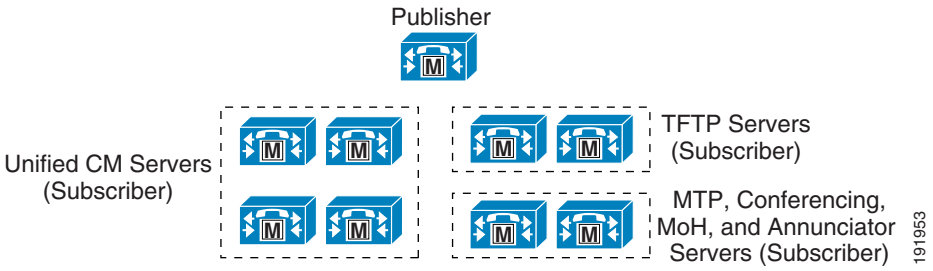
Within a Unified CM cluster, there are server nodes that provide unique services. Each of these services can coexist with others on the same server node. For example, in a small system it is possible to have a single server node providing database services, call processing services, and media resource services. As the scale and performance requirements of the cluster increase, many of these services should be moved to dedicated server nodes.

The following section describes the various functions performed by the server nodes that form a Unified CM cluster, and it provides guidelines for deploying the server nodes in ways that achieve the desired scale, performance, and resilience.

Cluster Server Nodes

Figure 9-1 illustrates a typical Unified CM cluster consisting of multiple server nodes. There are two types of Unified CM server nodes, publisher and subscriber. These terms are used to define the database relationship during installation.

Figure 9-1 Typical Unified CM Cluster



Publisher

The publisher is a required server node in all clusters, and as shown in Figure 9-1, there can be only one publisher per cluster. This server node is the first to be installed and provides the database services to all other subscribers in the cluster. The publisher node is the only server node that has full read and write access to the configuration database.

On larger systems with more than 1250 users, Cisco recommends a dedicated publisher to prevent administrative operations from affecting the telephony services. A dedicated publisher does not provide call processing or TFTP services running on the node. Instead, other subscriber nodes within the cluster provide these services.

The choice of the VM configuration for the publisher should be based on the desired scale and performance of the cluster. Cisco recommends that the publisher have the same server node performance capability as the call processing subscribers.

Subscriber

When the software is installed initially, only the database and network services are enabled. All subscriber nodes subscribe to the publisher to obtain a copy of the database information. However, in order to reduce initialization time for the Unified CM cluster, all subscriber nodes in the cluster attempt to use their local copy of the database when initializing. This reduces the overall initialization time for a Unified CM cluster. All subscriber nodes rely on change notification from the publisher or other subscriber nodes in order to keep their local copy of the database updated.

As shown in Figure 9-1, multiple subscriber nodes can be members of the same cluster. Subscriber nodes include Unified CM call processing subscriber nodes, TFTP subscriber nodes, and media resource subscriber nodes that provide functions such as conferencing and music on hold (MoH).

Call Processing Subscriber

A call processing subscriber is a server node that has the Cisco CallManager Service enabled. Once this service is enabled, the node is able to perform call processing functions. Devices such as phones, gateways, and media resources can register and make calls only to servers with this service enabled. As shown in Figure 9-1, multiple call processing subscribers can be members of the same cluster. In fact, Unified CM supports up to eight call processing subscriber nodes per cluster.

TFTP Subscriber

A TFTP subscriber or server node performs two main functions as part of the Unified CM cluster:

- The serving of files for services, including configuration files for devices such as phones and gateways, binary files for the upgrade of phones as well as some gateways, and various security files
- Generation of configuration and security files, which are usually signed and in some cases encrypted before being available for download

The Cisco TFTP service that provides this functionality can be enabled on any server node in the cluster. However, in a cluster with more than 1250 users, other services might be impacted by configuration changes that can cause the TFTP service to regenerate configuration files. Therefore, Cisco recommends that you dedicate a specific subscriber node to the TFTP service, as shown in Figure 9-1, for a cluster with more than 1250 users or any features that cause frequent configuration changes.

Cisco recommends that you use the same VM configuration for the TFTP subscribers as used for the call processing subscribers.

Media Resource Subscriber

A media resource subscriber or server node provides media services such as conferencing and music on hold to endpoints and gateways. These types of media resource services are provided by the Cisco IP Voice Media Streaming Application service, which can be enabled on any server node in the cluster.

Media resources include:

- Music on Hold (MoH) — Provides multicast or unicast music to devices that are placed on hold or temporary hold, transferred, or added to a conference. (See Music on Hold, page 7-17.)
- Annunciator service — Provides announcements in place of tones to indicate incorrectly dialed numbers or call routing unavailability. (See Annunciator, page 7-15.)
- Conference bridges — Provide software-based conferencing for instant and permanent conferences. (See Transcoding, page 7-5.)
- Media termination point (MTP) services — Provide features for H.323 clients, H.323 trunks, and Session Initiation Protocol (SIP) endpoints and trunks. (See Media Termination Point (MTP), page 7-7.)

Because of the additional processing and network requirements for media resource services, it is essential to follow all guidelines for running media resources within a cluster. Generally, Cisco recommends non-dedicated media resource subscribers for multicast MoH and annunciator services, but dedicated media resource subscribers as shown in Figure 9-1 are recommended for unicast MoH as well as large-scale software-based conferencing and MTPs unless those services are within the design guidelines detailed in the chapter on Media Resources, page 7-1.

Additional Cluster Services

In addition to the specific types of subscriber nodes within a Unified CM cluster, there are also other services that can be run on the Unified CM call processing subscriber nodes to provide additional functionality and enable additional features.

Computer Telephony Integration (CTI) Manager

The CTI Manager service acts as a broker between the Cisco CallManager service and TAPI or JTAPI integrated applications. This service is required in a cluster for any applications that utilize CTI. The CTI Manager service provides authentication of the CTI application and enables the application to monitor and/or control endpoint lines. CTI Manager can be enabled only on call processing subscribers, thus allowing for a maximum of eight nodes running the CTI Manager service in a cluster.

For more details on CTI Manager, see Computer Telephony Integration (CTI), page 9-28.

Unified CM Applications

Various types of application services can be enabled on Unified CM, such as Cisco Unified CM Assistant, Extension Mobility, and Web Dialer. For detailed design guidance on these applications, see the chapter on Cisco Unified CM Applications, page 18-1. The Cisco IM and Presence service can also be added (see the chapter on Collaboration Instant Messaging and Presence, page 20-1).

Mixing Unified CM VM Configurations

Mixing VM configurations within a Unified CM cluster is allowed, but Cisco recommends using the same VM configuration for all Unified CM nodes in a cluster. Cisco also recommends that the VM configuration used for the Unified CM publisher should not be smaller than any other Unified CM VM configuration used in the same cluster and that the VM configuration used for the backup subscribers should not be smaller than the VM configuration used for the primary subscribers.

When mixing VM configurations within a cluster, differences in capacity between the various VM configurations must be considered because the supported overall cluster capacity is limited by the cluster capacity corresponding to the smallest VM configuration within the cluster.

For example, if you mix one Unified CM call processing pair using the 7.5k VM configuration and two Unified CM call processing pairs using the 10k VM configuration, the overall cluster capacity that is supported corresponds to the cluster capacity of all nodes using the 7.5k VM configuration. With 3 call processing pairs in this example, the cluster capacity is limited to 22.5k endpoints (3 ∗ 7,500). To overcome this cluster capacity limitation, one option is to deploy separate clusters and connect those clusters with SIP trunks.

Mixing Hardware Platforms and Business Edition Platforms

Mixing different types of hardware platforms within a Unified CM cluster is also allowed, but because all VM configurations are not supported on all server hardware, this might result in mixing VM configurations and therefore might impact the overall cluster capacity. (See Mixing Unified CM VM Configurations, page 9-8, for details.) In addition to that, if Business Edition 6000 is part of the platform mix, the rules specific to the Business Edition 6000 solution must be taken into consideration.

Example 9-1 Mixing BE6000M and BE7000

Unified CM deployed as part of BE6000M is limited to 1,000 users and 1,200 devices, regardless of the number of cluster nodes. Adding other nodes is possible, whether or not they are part of BE7000 and whether or not those additional nodes are using larger VM configurations. It can provide redundancy and/or geographic distribution, but it does not increase the cluster capacity because of the BE6000 sizing rules. The Unified CM cluster capacity with BE6000M is still limited to 1,000 users and 1,200 devices when nodes are added. Similar restrictions apply with BE6000H, which is limited to a maximum of 1,000 users and 2,500 devices, regardless of the node count.

Example 9-2 Mixing Small TRC and BE7000

With the Small Tested Reference Configuration (TRC) that is not part of the Cisco Business Edition 6000M or 6000H solution, while the capacity of a node is limited to 1,000 users or devices, the capacity of the Unified CM cluster is not limited to 1,000 users or devices. If you add multiple nodes in a cluster, more than 1,000 users and devices can be supported. But on the Small TRC hardware platform, only the 1k-user VM configuration is supported. So if some Unified CM nodes running on the Small TRC and some running on BE7000 are mixed in the same cluster (as mentioned in the section on Mixing Unified CM VM Configurations, page 9-8), the overall cluster capacity that is supported is limited by the cluster capacity corresponding to the node using the smallest VM configuration – the 1,000-user VM configuration in this case. For example, if one Unified CM call processing pair is running on the Small TRC (with the 1k-user VM configuration) and another Unified CM call processing pair is running on the BE7000, the supported Unified CM cluster capacity is limited to 2,000 users and/or devices (2∗1,000), even if Unified CM VM configurations larger than 1k are deployed on BE7000.

Mixing servers from different vendors is allowed, but this would be under the specifications-based hardware policy, and Unified CM performance is not guaranteed on this type of platform mix.

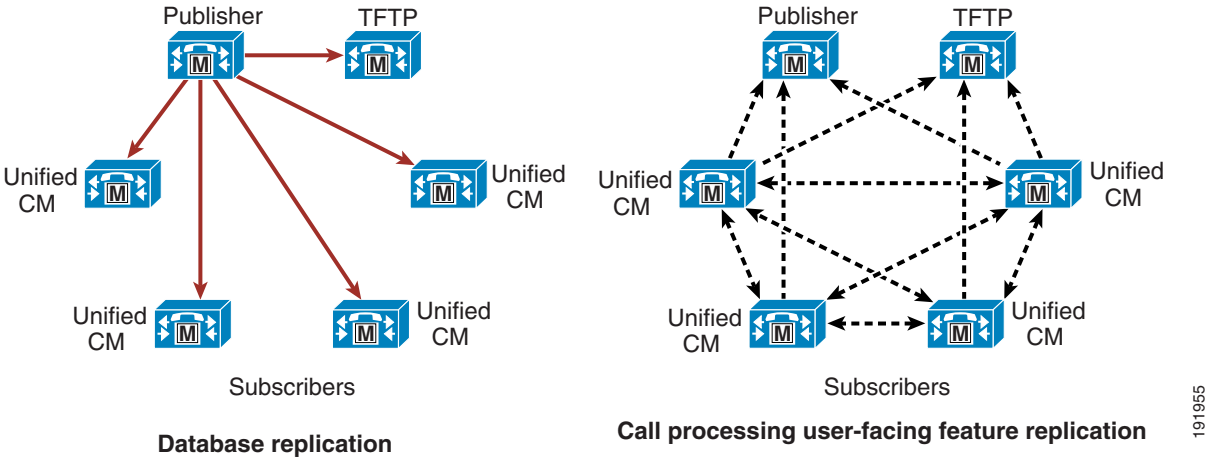
Intracluster Communications

There are two primary kinds of intracluster communications, or communications within a Unified CM cluster (see Figure 9-2 and Figure 9-3.) The first is a mechanism for distributing the database that contains all the device configuration information (see “Database replication” in Figure 9-2). The configuration database is stored on a publisher node, and a copy is replicated to the subscriber nodes of the cluster. Most of the database changes are made on the publisher and are then communicated to the subscriber databases, thus ensuring that the configuration is consistent across the members of the cluster and facilitating spatial redundancy of the database.

Database modifications for user-facing call processing features are made on the subscriber nodes to which an end-user device is registered. The subscriber nodes then replicate these database modifications to all the other nodes in the cluster, thus providing redundancy for the user-facing features. (See "Call processing user-facing feature replication" in Figure 9-2.) These features include:

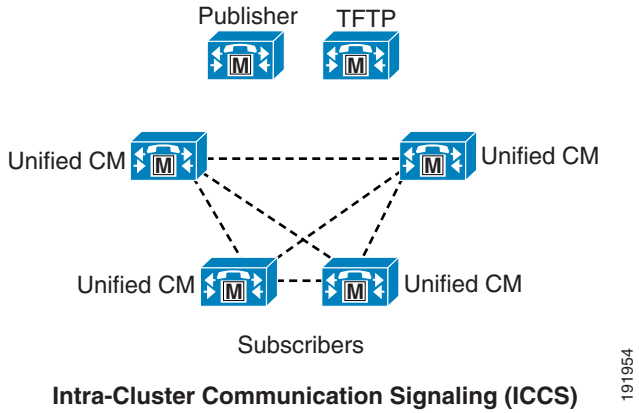
- Call Forward All (CFA)
- Message waiting indicator (MWI)
- Privacy Enable/Disable
- Extension Mobility login/logout
- Hunt Group login/logout
- Device Mobility
- Certificate Authority Proxy Function (CAPF) status for end users and applications users
- Credential hacking and authentication

Figure 9-2 Replication of the Database and User-Facing Features



The second type of intracluster communication, called Intra-Cluster Communication Signaling (ICCS), involves the propagation and replication of run-time data such as registration of devices, locations bandwidth, and shared media resources (see Figure 9-3). This information is shared across all members of a cluster running the Cisco CallManager Service (call processing subscribers), and it ensures the optimum routing of calls between members of the cluster and associated gateways.

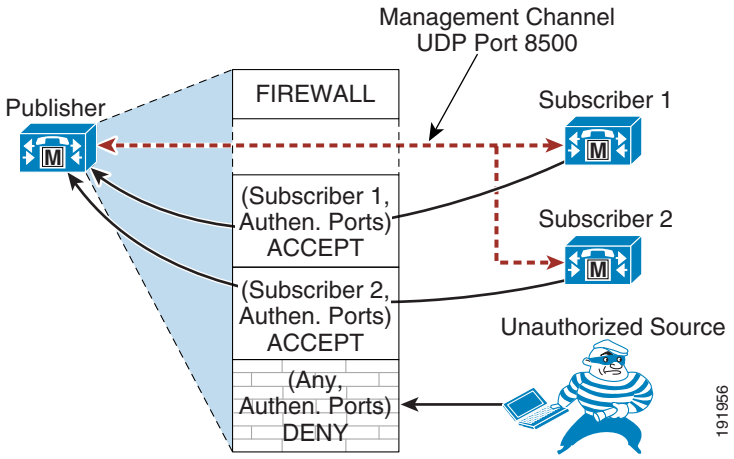
Figure 9-3 Intra-Cluster Communication Signaling (ICCS)



Intracluster Security

Each server node in a Unified CM cluster runs an internal dynamic firewall. The application ports on Unified CM are protected by source IP filtering. The dynamic firewall opens these application ports only to authenticated or trusted server nodes. (See Figure 9-4.)

Figure 9-4 Intracluster Security



This security mechanism is applicable only between server nodes in a single Unified CM cluster. Unified CM subscribers are authenticated in a cluster before they can access the publisher's database. The intra-cluster communication and database replication take place only between authenticated server nodes. During the installation process, a subscriber node is authenticated to the publisher using a pre-shared key authentication mechanism. The authentication process involves the following steps:

1. Install the publisher node using a security password.
2. Configure the subscriber node on the publisher by using Unified CM Administration.
3. Install the subscriber node using the same security password used during publisher server installation.
4. After the subscriber is installed, the server node attempts to establish connection to the publisher on a management channel using UDP 8500. The subscriber sends all the credentials to the publisher, such as hostname, IP address, and so forth. The credentials are authenticated using the security password used during the installation process.
5. The publisher verifies the subscriber's credentials using its own security password.
6. The publisher adds the subscriber as a trusted source to its dynamic firewall table if the information is valid. The subscriber is allowed access to the database.
7. The subscriber gets a list of other subscriber nodes from the publisher. All the subscribers establish a management channel with each other, thus creating a mesh topology.

General Clustering Guidelines

The following guidelines apply to all Unified CM clusters:

- Cisco recommends using the same VM configuration for all nodes in a cluster. Mixing Unified CM VM configurations is allowed, but there are design implications and limits. For more details, refer to the section on Mixing Unified CM VM Configurations, page 9-8.
- Under normal circumstances, place all members of the cluster within the same LAN or MAN.
- If the cluster spans an IP WAN, follow the guidelines for clustering over an IP WAN as specified in the section on Clustering Over the IP WAN, page 10-43.
- A Unified CM cluster may contain as many as 20 server nodes, of which a maximum of eight call processing subscribers (nodes running the Cisco CallManager Service) are allowed. The other server nodes within the cluster may be configured as a dedicated database publisher, dedicated TFTP subscriber, or media resource subscriber.
- When deploying a two-node cluster, Cisco recommends that you do not exceed 1250 users in the cluster. Above 1250 users, a dedicated publisher and separate server nodes for primary and backup call processing subscribers is recommended.
- Business Edition 6000 provides a single instance of Unified CM (a Unified CM publisher that also handles call processing). Additional Business Edition 6000 server(s) may be deployed to provide subscriber redundancy either in an active/standby or load balancing fashion for Unified CM as well as some other co-resident applications. However, adding new nodes and new hardware platforms does not increase capacity. For example, the user and device capacities do not increase.
- Each Unified CM node instance can be a publisher node, call processing subscriber node, TFTP subscriber node, or media resource subscriber node. Only a single publisher node per cluster is supported.
- With virtualization, Unified CM no longer supports the Cisco Messaging Interface (CMI) service for Simplified Message Desk Interface (SMDI) integrations, fixed MoH audio source integration for live MoH audio feeds using the audio cards (MOH-USB-AUDIO=), or flash drives to these servers. The following alternate options are available:
  - For MoH live audio source feed, consider using Cisco IOS-based gateway multicast MoH for live audio source connectivity.
  - For saving system install logs, use virtual floppy softmedia.
  - There is no alternate option for the Cisco Messaging Interface (CMI) service for Simplified Message Desk Interface (SMDI) integrations.

High Availability for Call Processing

You should deploy the call processing services within a Unified Communications System in a highly available manner so that a failure of a single call processing component will not render all call processing services unavailable.

Hardware Platform High Availability

You should select the call processing platform based not only on the size and scalability of a particular deployment, but also on the redundant nature of the platform hardware.

When possible, choose platforms with dual power supplies to ensure that a single power supply failure will not result in the loss of a platform. Plug platforms with dual power supplies into two different power sources to avoid the failure of one power circuit causing the entire platform to fail. The use of dual power supplies combined with the use of uninterruptible power supply (UPS) sources will ensure maximum power availability. In deployments where dual power supply platforms are not feasible, Cisco still recommends the use of a UPS in situations where building power does not have the required level of power availability.

Providing hardware platform high availability is even more critical when deploying virtualization because a platform failure could result in the failure of all the virtual machines running on that hardware platform. When possible, avoid running multiple instances of the same application that have similar functions on the same physical server; instead, distribute those virtual machines across multiple servers and even across multiple chassis if possible when using Cisco UCS B-Series Blade Servers.

Network Connectivity High Availability

Connectivity to the IP network is also a critical consideration for maximum performance and high availability. With Cisco Unified CME, use a minimum of two ports to connect to the network. With Unified CM, high availability for the network connectivity is attained at the host level by configuring the hypervisor virtual switch with multiple uplinks and thus by using multiple physical ports on the hardware platform. Therefore, a single virtual NIC defined in the OVA setting is sufficient. If you are using the VMware vSphere virtual switch, for example, configure NIC teaming for the switch uplinks. Also connect those multiple ports to a minimum of two upstream switches to provide resiliency if an upstream switch fails.

Connect platforms to the network at the highest possible speed to ensure maximum throughput, typically 1 Gbps or even 10 Gbps when using the UCS B-Series platform. Ensure that platforms are connected to the network using full-duplex.

In addition to speed and duplex of IP network connectivity, equally important is the resilience of this network connectivity. Unified communications deployments are highly dependent on the underlying network connectivity for true redundancy. For this reason it is critical to deploy and configure the underlying network infrastructure in a highly resilient manner. For details on designing highly available network infrastructures, see the chapter on Network Infrastructure, page 3-1. In all cases, the network should be designed so that, given a switch or router failure within the infrastructure, a majority of users will have access to a majority of the services provided within the deployment.

To maximize call processing availability, locate and connect call processing platforms in separate buildings and/or separate network switches when possible to ensure that the impact to call processing will be minimized if there is a failure of the building or network infrastructure switch. With Unified CM

call processing, this means distributing cluster server nodes among multiple buildings or locations within the LAN or MAN deployment whenever possible. And at the very least, it means physically distributing network connections between different physical network switches in the same location.

Furthermore, even though Cisco Unified CME is a standalone call processing entity, providing physical distribution and therefore redundancy for this call processing entity still makes sense when deploying multiple call processing entities. Whenever possible in those scenarios, install each instance of Unified CME in a different physical location within the network, or at the very least physically attach them to different network switches.

Unified CM High Availability

Because of the underlying Unified CM clustering mechanism, a Unified Communications System has additional high availability considerations above and beyond hardware platform disk and power component redundancy, physical network location, and connectivity redundancy. This section examines call processing subscriber redundancy considerations, call processing load balancing, and redundancy of additional cluster services.

Call Processing Redundancy

Unified CM provides the following call processing redundancy configuration options or schemes:

- Two to one (2:1) — For every two primary call processing subscribers, there is one shared secondary or backup call processing subscriber.
- One to one (1:1) — For every primary call processing subscriber, there is a secondary or backup call processing subscriber.

These redundancy schemes are facilitated by the built-in registration failover mechanism within the Unified CM cluster architecture, which enables endpoints to re-register to a backup call processing subscriber node when the endpoint's primary call processing subscriber node fails. The registration failover mechanism can achieve failover rates for Skinny Client Control Protocol (SCCP) IP phones of approximately 125 registrations per second. The registration failover rate for Session Initiation Protocol (SIP) phones is approximately 40 registrations per second.

The call processing redundancy scheme you select determines not only the fault tolerance of the deployment, but also the fault tolerance of any upgrade.

With 1:1 redundancy, multiple primary call processing subscriber failures can occur without impacting call processing capabilities. With 2:1 redundancy, on the other hand, only one of the primary call processing subscribers out of the two primary call processing subscribers that share a backup call processing subscriber can fail without impacting call processing. However, if the total number of endpoints registered across both primary subscribers and the traffic to those two primary subscribers are within the capacity limits of the backup subscriber, then the backup subscriber is able to handle the failure of both primary subscribers.

Note Do not deploy 2:1 redundancy if the total capacity utilization across the two primary subscribers would exceed the capacity of the backup subscriber. For example, if the call processing capacity or endpoints capacity utilization exceeds 50% on both primary subscribers, the backup subscriber would not be able to handle call processing services properly if both primary subscribers fail. In these scenarios, for example, some endpoints might not be able to register, some new calls might not be established, and some services and features might not operate properly because the backup subscriber system capacity has been exceeded.

Likewise, with the 1:1 redundancy scheme, upgrades to the cluster can be performed with only a single set of endpoint registration failover periods impacting the call processing services. Whereas with the 2:1 redundancy scheme, upgrades to the cluster can require multiple registration failover periods.

A Unified CM cluster can be upgraded with minimal impact to the services. Two different versions (releases) of Unified CM may be on the same server node, one in the active partition and the other in the inactive partition. All services and devices use the Unified CM version in the active partition for all Unified CM functionality. During the upgrade process, the cluster operations continue using its current release of Unified CM in the active partition, while the upgrade version gets installed in the inactive partition. Once the upgrade process is complete, the server nodes can be rebooted to switch the inactive partition to the active partition, thus running the new version of Unified CM.

With the 1:1 redundancy scheme, the following steps enable you to upgrade the cluster while minimizing downtime:

Step 1 Install the new version of Unified CM in the inactive partition, first on the publisher and then on all subscribers (call processing, TFTP, and media resource subscribers). Do not reboot.

Step 2 Reboot the publisher and switch to the new version.

Step 3 Reboot the TFTP subscriber node(s) one at a time and switch to the new version.

Step 4 Reboot any dedicated media resource subscriber nodes one at a time and switch to the new version.

Step 5 Reboot the backup call processing subscribers one at a time and switch to the new version.

Step 6 Reboot the primary call processing subscribers one at a time and switch to the new version. Device registrations will fail-over to the previously upgraded and rebooted backup call processing subscribers. After each primary call processing subscriber is rebooted, devices will begin to re-register to the primary call processing subscriber.

With this upgrade method, there is no period (except for the registration failover period) when devices are registered to subscriber nodes that are running different versions of the Unified CM software. All these steps can be automated using Cisco Prime Collaboration.

While the 2:1 redundancy scheme allows for fewer server nodes in a cluster, registration failover occurs more frequently during upgrades, increasing the overall duration of the upgrade as well as the amount of time call processing services for a particular endpoint will be unavailable. Because there is only a single backup call processing subscriber per pair of primary call processing subscribers, it might be possible to reboot to the new version on only one of the primary call processing subscribers in a pair at a time in order to prevent oversubscribing the single backup call processing subscriber. As a result, there may be a period of time after the first primary call processing subscriber in each pair is switched to the new version, in which endpoint registrations will have to be moved from the backup subscriber to the newly upgraded primary subscriber before the endpoint registrations on the second primary subscriber can be moved to the backup subscriber to allow a reboot to the new version. During this time, not only will endpoints on the second primary call processing subscriber be unavailable while they re-register to the backup subscriber, but until they re-register to a node running the new version, they will also be unable to reach endpoints on other subscriber nodes that have already been upgraded.

Note Before you do an upgrade, Cisco recommends that you back up the Unified CM and Call Detail Record (CDR) database to an external network directory using the Disaster Recovery Framework. This practice will prevent any loss of data if the upgrade fails.

Note Because an upgrade of a Unified CM cluster results in a period of time in which some or most devices lose registration and call processing services temporarily, you should plan upgrades in advance and implement them during a scheduled maintenance window. While downtime and loss of services to devices can be minimized by selecting the 1:1 redundancy scheme, there will still be some period of time in which call processing services are not available to some or all users.

For more information on upgrading Unified CM, refer to the installation and upgrade guides available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-guides-list.html

Unified CM Redundancy with Survivable Remote Site Telephony (SRST)

Cisco IOS SRST provides highly available call processing services for endpoints in locations remote from the Unified CM cluster. Unified CM clustering redundancy schemes certainly provide a high level of redundancy for call processing and other application services within a LAN or MAN environment. However, for remote locations separated from the central Unified CM cluster by a WAN or other low-speed links, SRST can be used as a redundancy method to provide basic call processing services to these remote locations in the event of loss of network connectivity between the remote and central sites. Cisco recommends deploying SRST-capable Cisco IOS routers at each remote site where call processing services are considered critical and need to be maintained in the event that connectivity to the Unified CM cluster is lost. Endpoints at these remote locations must be configured with an appropriate SRST reference within Unified CM so that the endpoint knows what address to use to connect to the SRST router for call processing services when connectivity to Unified CM subscribers is unavailable.

Cisco Unified Enhanced SRST (E-SRST) on a Cisco IOS router can also be used at a remote site to provide backup call processing functionality in the event that connectivity to the central Unified CM cluster is lost. E-SRST provides more telephony features for the IP phones than are available with the regular SRST feature on a router. However, the endpoint capacities for Unified E-SRST are typically less than for basic SRST. Both SRST and E-SRST are supported with Cisco Unified SRST Manager, which synchronizes configurations from Unified CM with SRST and E-SRST, thus reducing manual configuration required in the branch SRST or E-SRST router and enabling users to have a similar calling experience in both SRST and normal modes.

Call Processing Subscriber Redundancy

Depending on the redundancy scheme chosen (see Call Processing Redundancy, page 9-14), the call processing subscriber will be either a primary (active) subscriber or a backup (standby) subscriber. In the load-balancing option, the subscriber can be both a primary and backup subscriber. When planning the design of a cluster, you should generally dedicate the call processing subscribers to this function. In larger-scale or higher-performance clusters, the call processing service should not be enabled on the publisher and TFTP subscriber nodes. 1:1 redundancy uses dedicated pairs of primary and backup subscribers, while 2:1 redundancy uses a pair of primary subscribers that share one backup subscriber.

The following figures illustrate typical cluster configurations to provide call processing redundancy with Unified CM.

Figure 9-5 Basic Redundancy Schemes

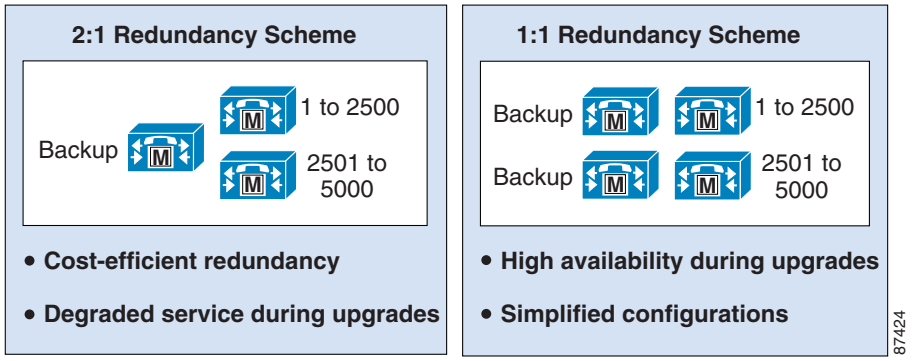


Figure 9-5 illustrates the two basic redundancy schemes available. In each case the backup server node must be capable of handling the capacity of at least a single primary call processing server node failure. In the 2:1 redundancy scheme, the backup might have to be capable of handling the failure of a single call processing server node or potentially both primary call processing server nodes, depending on the requirements of a particular deployment. For information on capacity sizing and choosing the VM configurations, see the section on Capacity Planning for Call Processing, page 9-23.

Note 2:1 redundancy is not supported with the 10,000-User VM configuration due to potential overload on the backup subscriber.

Figure 9-6 1:1 Redundancy Configuration Options

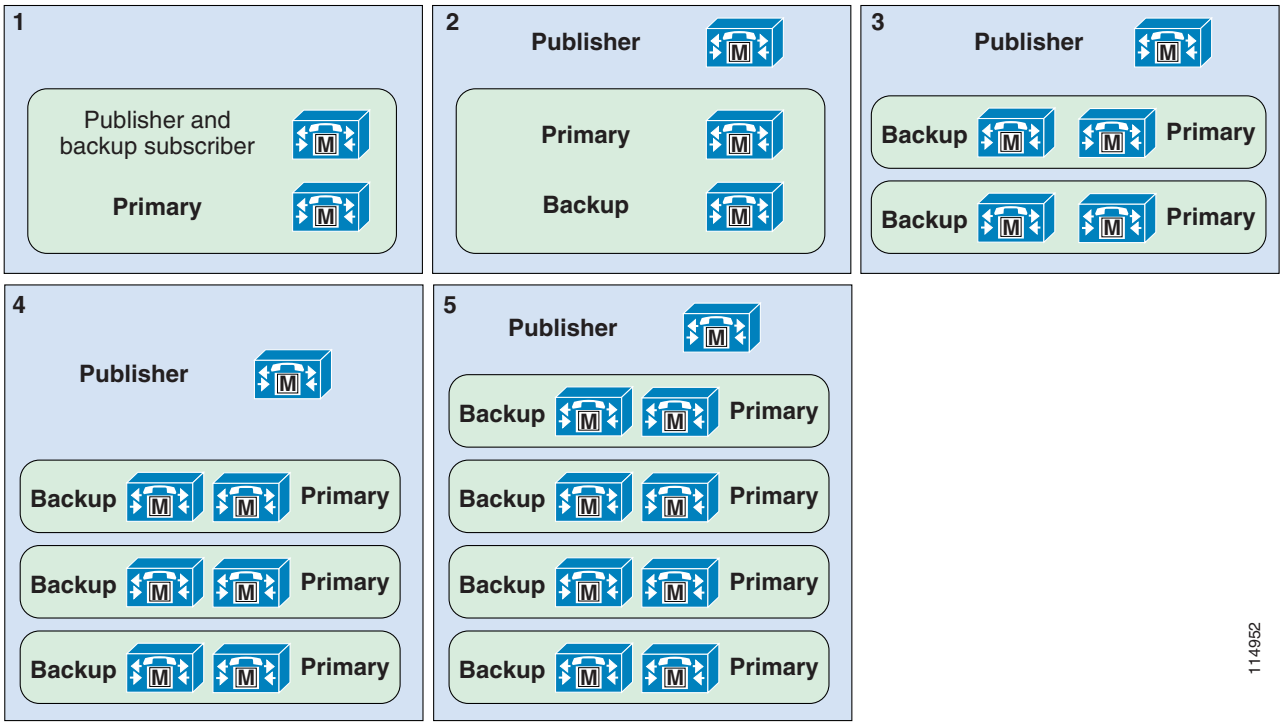
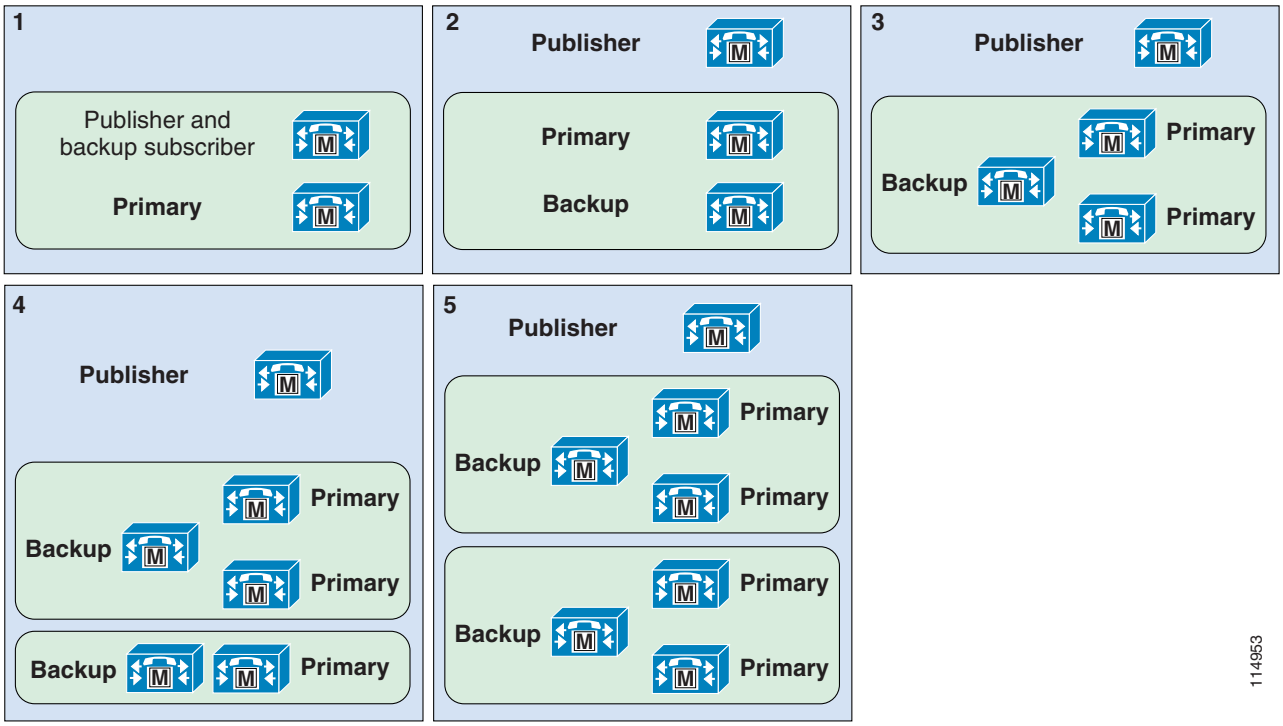


Figure 9-7 2:1 Redundancy Configuration Options



In Figure 9-6, the five options shown all indicate 1:1 redundancy. In Figure 9-7, the five options shown all indicate 2:1 redundancy. In both cases, Option 1 is used for clusters supporting less than 1,250 users and includes Unified CM deployments with Cisco Business Edition 6000. Options 2 through 5 illustrate increasingly scalable clusters for each redundancy scheme. The exact scale depends on the hardware platforms chosen or required.

These illustrations show only publisher and call processing subscribers. They do not account for other subscriber nodes such as TFTP and media resources.

Note It is possible to define up to three call processing subscribers per Unified CM group. Adding a tertiary subscriber for additional backup extends the above redundancy schemes to 2:1:1 or 1:1:1 redundancy. However, with the exception of using tertiary subscriber nodes in deployments with clustering over the WAN (see Remote Failover Deployment Model, page 10-54), tertiary subscriber redundancy is not recommended for endpoint devices located in remote sites because failover to SRST will be further delayed if the endpoint must check for connectivity to a tertiary subscriber. The tertiary subscribers also count against the maximum number of call processing subscribers in a cluster (8 call processing subscriber nodes).

Although not shown in the Figure 9-6 or Figure 9-7, it is also possible to deploy a single-node cluster. The single-node cluster should not exceed 1000 endpoint configuration and registrations. Note that in a single-node configuration, there is no backup call processing subscriber and therefore no cluster redundancy mechanism. Survivable Remote Site Telephony (SRST) can be used as a redundancy mechanism in these types of deployments to provide minimal call processing services during periods when Unified CM is not available.

Load Balancing

In Unified CM clusters with the 1:1 redundancy scheme, device registration and call processing services can be load-balanced across the primary and backup call processing subscriber.

Normally a backup server node has no devices registered to it unless its primary is unavailable. This makes it easier to troubleshoot a deployment because there is a maximum of four primary call processing subscriber nodes that will be handling the call processing load at a given time. Further, this potentially simplifies configuration by reducing the number of Unified CM redundancy groups and device pools.

In a load-balanced deployment, up to half of the device registration and call processing load can be moved from the primary to the secondary subscriber by using the Unified CM redundancy groups and device pool settings. In this way each primary and backup call processing subscriber pair provides device registration and call processing services to as many as half of the total devices serviced by this pair of call processing subscribers. This is referred to as 50/50 load balancing. The 50/50 load balancing model provides the following benefits:

- Load sharing — The registration and call processing load is distributed on multiple server nodes, which can provide faster response time.
- Faster failover and failback — Because all devices (such as IP phones, CTI ports, gateways, trunks, voicemail ports, and so forth) are distributed across all active subscribers, only some of the devices fail-over to the secondary subscriber if the primary subscriber fails. In this way, you can reduce by 50% the impact of any server node becoming unavailable.

To plan for 50/50 load balancing, calculate the capacity of a cluster without load balancing, and then distribute the load across the primary and backup subscribers based on devices and call volume. To allow for failure of the primary or the backup server node, do not let the total load on the primary and secondary subscribers exceed that of a single subscriber node.

Note During upgrades of a Unified CM cluster with 50/50 load balancing, upgrades to the backup call processing subscriber will result in devices registered to that subscriber (half of the total devices serviced by the primary and backup subscriber pair) failing over to the primary call processing subscriber.

TFTP Redundancy

Cisco recommends deploying more than one dedicated TFTP subscriber node for a large Unified CM cluster, thus providing redundancy for TFTP services. While two TFTP subscribers are typically sufficient, more than two TFTP server nodes can be deployed in a cluster.

In addition to providing one or more redundant TFTP subscribers, you must configure endpoints to take advantage of these redundant TFTP nodes. When configuring the TFTP options using DHCP or statically, define a TFTP subscriber node IP address array containing the IP addresses of both TFTP subscriber nodes within the cluster. In this way, by creating two DHCP scopes with two different IP address arrays (or by manually configuring endpoints with two different TFTP subscriber node IP addresses), you can assign half of the endpoint devices to use TFTP subscriber A as the primary and TFTP subscriber B as the backup, and the other half to use TFTP subscriber B as the primary and TFTP subscriber A as the backup. In addition to providing redundancy during a failure of one TFTP subscriber, this method of distributing endpoints across multiple TFTP subscribers provides load balancing so that one TFTP subscriber is not handling all the TFTP service load.

Note When adding a specific binary or firmware load for a phone or gateway, you must add the file(s) to each TFTP subscriber node in the cluster.

CTI Manager Redundancy

All CTI integrated applications communicate with a call processing subscriber node running the CTI Manager service. Further, most CTI applications have the ability to specify redundant CTI Manager service nodes. For this reason, Cisco recommends activating the CTI Manager service on at least two call processing subscribers within the cluster. With both a primary and backup CTI Manager configured, in the event of a failure the application will switch to a backup CTI Manager to receive CTI services.

As stated previously, the CTI Manager service can be enabled only on call processing subscribers, therefore there is a maximum of eight CTI Managers per cluster. Cisco recommends that you load-balance CTI applications across the enabled CTI Managers in the cluster to provide maximum resilience, performance, and redundancy.

Generally, it is good practice to associate devices that will be controlled or monitored by a CTI application with the same server node pair used for the CTI Manager service. For example, an interactive voice response (IVR) application requires four CTI ports. They would be provisioned as follows, assuming the use of 1:1 redundancy and 50/50 load balancing:

- Two CTI ports would have a Unified CM redundancy group of server node A as the primary call processing subscriber and server node B as the backup subscriber. The other two ports would have a Unified CM redundancy group of server node B as the primary subscriber and server node A as the backup subscriber.
- The IVR application would be configured to use the CTI Manager on subscriber A as the primary and subscriber B as the backup.

The above example allows for redundancy in case of failure of the CTI Manager on subscriber A and also allows for the IVR call load to be spread across two server nodes. This approach also minimizes the impact of a Unified CM subscriber node failure.

For more details on CTI and CTI Manager, see Computer Telephony Integration (CTI), page 9-28.

Virtual Machine Placement and Hardware Platform Redundancy

With virtualization there are redundancy considerations because of the virtual nature of server nodes: namely, the installation and residency of Unified CM server node instances across physical servers.

As illustrated by the example in Figure 9-8, observe the following guidelines when deploying Unified CM to ensure the highest level of call processing redundancy:

- Each primary call processing subscriber node instance should reside on a different physical server than its backup call processing subscriber node instance. This ensures that the failure of a server containing the primary call processing node instance does not impact the system's ability to provide endpoints with access to their backup call processing subscriber node.
- When deploying multiple TFTP or media resource subscriber nodes instances for redundancy of those services, always distribute redundant subscriber nodes across more than one server to ensure that a failure of a single server does not eliminate those services. This ensures that, given the failure of a blade containing a TFTP or media resource subscriber, endpoints will still be able to access TFTP and media resource services on a subscriber node residing on another server. Endpoints can also be distributed among redundant TFTP and media resource subscriber node instances to balance system load in non-failure scenarios.
- When deploying CTI applications, always make sure that call processing subscriber node instances running the CTI Manager service are distributed across more than one server to ensure that a failure of a single server does not eliminate CTI services. Further, CTI applications should be configured to use the CTI Manager service running on the subscriber node instance on one server as the primary CTI Manager and the CTI Manager service running on the subscriber node on another server as the backup CTI Manager.

Figure 9-8 Unified CM Server Node Distribution on UCS

When using blade servers with a chassis (for example, B-Series blade servers with a Cisco UCS 5100 Blade chassis), in addition to distributing subscriber node instances across multiple blades, you may distribute subscriber node instances across multiple blade chassis for additional redundancy and scalability.

For more information about redundancy and provisioning of host resources for virtual machines, refer to the documentation at https://www.cisco.com/go/virtualized-collaboration.

Cisco Business Edition High Availability

With Cisco Business Edition 6000M, Cisco Business Edition 6000H, and Cisco Business Edition 7000, high availability is provided by clustering additional Cisco Unified CM node(s). Additional Business Edition server(s) can be deployed to provide high availability for call processing as well as other applications and services.

Note More than two physical servers may be clustered to provide additional redundancy and/or geographic distribution as with a clustering over the WAN deployment. However, with Cisco Business Edition 6000, the additional server(s) only provides redundancy and not a capacity increase. For example, with BE6000M and BE6000H, the total number of users across the cluster may not exceed 1,000. A deployment exceeding this limit is considered to be a standard Unified CM cluster, and as such the deployment must follow high availability design guidance for standard Unified CM. (See Unified CM High Availability, page 9-14.) With Cisco Business Edition 7000, the capacity is not limited to 1,000 users; rather, the standard application capacity planning and design rules apply.

Capacity Planning for Call Processing

Call processing capacity planning is critical for successful unified communications deployments. This section discusses capacity planning for Cisco Unified CM, whether or not it is part of the Cisco Business Edition 6000 or 7000 solution. It also covers Cisco Business Edition 4000 and Unified CME.

Unified CM Capacity Planning

Unified CM capacity depends on the hardware platform, the VM configuration, and the deployment requirements. It also depends on whether or not Unified CM is deployed as part of Cisco Business Edition 6000. Table 9-2 lists some general Unified CM capacity limits.

Table 9-2 Cisco Unified CM Capacity Limits

Capacity Information	Cisco Business Edition 6000M	Cisco Business Edition 6000H	Cisco Business Edition 7000 and Enterprise Cisco Unified CM
Maximum number of users	1,000 per cluster	1,000 per cluster	Up to 10,000 per node; up to 40,000 per cluster[1]
Maximum number of endpoints	1,200 per cluster	2,500 per cluster	Up to 10,000 per node; up to 40,000 per cluster[1]
How to perform capacity planning	Capacity information in product documentation[2]	Capacity information in product documentation[2]	Product documentation, SRND guidelines, and Cisco Collaboration Sizing Tool[3]

1. Could be higher with a Megacluster deployment.
2. When necessary, capacity planning can be based on the Cisco Collaboration Sizing Tool instead of the fixed capacity numbers in the Business Edition product documentation. However, the Unified CM cluster is still limited to 1,000 users.
3. For deployments with up to 10,000 users or endpoints (whichever limit is reached first), capacity planning can be done with the simplified sizing available in the Cisco Preferred Architecture for Enterprise Collaboration.

Cisco Business Edition 6000M/H Capacity Planning

With the Cisco Business Edition 6000, Unified CM is deployed with a specific VM configuration and Unified CM capacity is fixed. Cisco Unified CM capacity planning is simple and does not rely on the Cisco Collaboration Sizing Tool.

Unified CM deployed with Cisco Business Edition 6000M supports up to 1,000 users, 1,200 devices, and 5,000 BHCA. With Business Edition 6000H, up to 1,000 users, 2,500 devices and 5,000 BHCA are supported.

With BE6000, adding nodes or hardware platforms is supported to provide high availability, but that does not increase capacity. The VM configurations specific to the Business Edition 6000 must be used. The larger VM configurations for 2,500, 7,500, or 10,000 users cannot be used with Business Edition 6000.

Unified CM deployed as part of Business Edition 6000 has some additional restrictions. For example, up to 50 sites and up to 100 contact center agents are supported with Cisco Business Edition 6000M/H. (For more details, refer to the Cisco Business Edition 6000 product documentation available at https://www.cisco.com/go/be6000.) If those requirements cannot be met but the number of users is still under 1,000, it is possible to take an alternate approach to the capacity planning with Cisco Business Edition 6000M/H. Instead of relying on fixed capacity specific to Business Edition 6000, the sizing could be done the same way as with Business Edition 7000 and enterprise Unified CM, based on the sizing guidelines in the product documentation, this SRND, and the Cisco Collaboration Sizing Tool. The 1,000-user VM configuration would have to be selected when using the Sizing Tool.

Cisco Business Edition 7000M/H and Cisco Unified CM Capacity Planning

With Cisco Business Edition 7000 or the enterprise (non-Business-Edition) version of Cisco Unified CM, various VM configurations with different corresponding capacities are available, and the capacity increases when nodes are added. Capacity planning is done based on the guidelines in this document and the Cisco Collaboration Sizing Tool. However, there is a simplified capacity planning method described in the *Cisco Preferred Architecture for Enterprise Collaboration CVD* (available at https://www.cisco.com/go/pa). This simplified sizing method can be used if the Unified CM deployment has less than 10,000 users or devices (whichever limit is reached first) and if specific sizing assumptions are met. If those sizing assumptions cannot be met, then simplified capacity planning cannot be used, and normal capacity planning based on the guidelines in the product documentation, this SRND, and the Cisco Collaboration Sizing Tool must be done instead. The sizing tool takes into account many more parameters; for example, the type of phone (SCCP, SIP, or mobile) and the phone security mode are taken into consideration. It is a more complex sizing process, but it can be customized to your specific deployment.

Enabling and increasing utilization of some Unified CM functions can have an impact on the call processing capabilities of the system and in some cases can reduce the overall capacity. These functions include tracing, call detail recording, highly complex dial plans, and other services that are co-resident on the Unified CM platform. Highly complex dial plans can include multiple line appearances as well as large numbers of partitions, calling search spaces, route patterns, translations, route groups, hunt groups, pickup groups, route lists, call forwarding, co-resident services, and other co-resident applications. All of these functions can consume additional resources within the Unified CM system.

You can use the following technique to improve system performance:

A Unified CM cluster with a very large dial plan containing many gateways, route patterns, translation patterns, and partitions, can take an extended amount of time to initialize when the Cisco CallManager Service is first started. If the system does not initialize within the default time, you can modify the system initialization timer (a Unified CM service parameter) to allow additional time for the configuration to initialize. For details on the system initialization time, refer to the online help for *Service Parameters* in Unified CM Administration.

Unified CM Capacity Planning Guidelines and Endpoint Limits

The following capacity guidelines apply to Cisco Unified CM as part of Cisco Business Edition 7000 or outside of Business Edition:

- Within a cluster, a maximum of 8 call processing subscriber nodes can be enabled with the Cisco CallManager Service. Other server nodes may be used for more dedicated functions such as publisher, TFTP subscribers, and media resources subscribers.
- Each Unified CM node can support registration for a maximum of 10,000 secured or unsecured SCCP or SIP endpoints. Each cluster can support configuration and registration for a maximum of 40,000 secured or unsecured SCCP or SIP endpoints.
- There are several VM configuration options for Cisco Unified CM available in the OVA, depending on the required capacity. The names of the VM configurations correspond to the maximum number of users per node, assuming that each user has one phone. When the ratio of number phones per user is different than one, the VM configuration names actually correspond to the maximum number of endpoints per node. Depending on different variables such as BHCA and feature set used, the actual number of users or endpoints could be lower. To validate the sizing of a deployment, use the Cisco Collaboration Sizing Tool, available at https://www.cisco.com/go/cst.

- The default trace setting for the CallManager service is 1,500 files of 10 MB for Signaling Distribution Layer (SDL) traces. Unless specific troubleshooting under high call rates requires increasing the maximum file setting, the default settings are sufficient for collecting sufficient traces in most circumstances.

For more information about Unified CM capacity planning considerations, including sizing limits as well as a complete discussion of system sizing, capacity planning, and deployment considerations, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Megacluster

The term *megacluster* defines and identifies certain Unified CM deployments that allow for further increases in scalability. A megacluster provides more device capacity through the support of additional Unified CM subscriber nodes, with a maximum of eight Unified CM subscriber pairs (1:1 redundancy) per megacluster, thus allowing for a maximum of 80,000 devices.

A megacluster can also be deployed where customers simply require non-locally redundant call processing functionality, rather than using Survivable Remote Site Telephony (SRST), to scale beyond the maximum eight sites allowed in a standard cluster deployment and up to 16 Unified CM subscriber nodes per megacluster. For example, consider a large hospital that has twelve locations and each location has only 1,000 devices. This total of 12,000 devices could be accommodated within a standard cluster, which has a maximum device capacity of 40,000 devices. However, in this case it is the need for additional Unified CM subscribers, rather than additional device capacity, that requires a megacluster deployment. In this example, a Unified CM subscriber node could be deployed in each location, and each Unified CM subscriber could serve as the primary subscriber for the local endpoints and as a backup subscriber for endpoints from another location.

When considering a megacluster deployment, the primary areas impacting capacity are as follows:

- The megacluster may contain a total of 21 server nodes consisting of 16 subscriber nodes, 2 TFTP server nodes, 2 music on hold (MoH) server nodes, and 1 publisher node.
- Unified CM must be deployed with the 7,500-user or 10,000-user VM configuration options.
- Redundancy model must be 1:1.

All other capacities relating to a standard cluster also apply to a megacluster. Note that support for a megacluster deployment is granted only following the successful review of a detailed design, including the submission of the results from the Cisco Collaboration Sizing Tool. For more information about the Cisco Collaboration Sizing Tool and the sizing of Unified CM standard clusters and megaclusters, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Due to the many potential complexities surrounding megacluster deployments, customers who wish to pursue such a deployment must engage either their Cisco Account Team, Cisco Advanced Services, or their certified Cisco Unified Communications Partner.

Note Unless otherwise specified, all information contained within this SRND that relates to call processing deployments (including capacity, high availability, and general design considerations) applies only to a standard cluster.

For more information about call processing sizing and for a complete discussion of system sizing, capacity planning, and deployment considerations, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Cisco Business Edition 4000 Capacity Planning

Cisco Business Edition 4000 supports a maximum of 200 endpoints. For more information, refer to the chapter on Collaboration Solution Sizing Guidance, page 25-1, and to the Business Edition 4000 documentation available at:

https://www.cisco.com/c/en/us/products/unified-communications/business-edition-4000/index.html

Unified CME Capacity Planning

When deploying Unified CME, it is critical to select a Cisco IOS router platform that provides the desired capacity in terms of number of supported endpoints required. In addition, platform memory capacity should also be considered if the Unified CME router is providing additional services above and beyond call processing, such as IP routing, DNS lookup, dynamic host configuration protocol (DHCP) address services, or VXML scripting.

Unified CME can support a maximum of 450 endpoints on a single Cisco IOS platform; however, each router platform has a different endpoint capacity based on the size of the system. Because Unified CME is not supported within the Cisco Collaboration Sizing Tool, it is imperative to follow capacity information provided in the Unified CME product data sheets available at

https://www.cisco.com/c/en/us/products/unified-communications/unified-communications-manager-express/datasheet-listing.html

Design Considerations for Call Processing

Observe the following design recommendations and guidelines when deploying Cisco call processing:

Cisco Unified CM

- Cisco Unified CM runs only as a virtualized application on the VMware Hypervisor. It does not run directly on a hardware platform without the VMware Hypervisor.
- You can enable a maximum of 8 call processing subscriber nodes (nodes running the Cisco CallManager Service) within a Cisco Unified CM cluster. Additional server nodes may be dedicated and used for publisher, TFTP, and media resources services. An approved megacluster deployment supports a maximum of 16 call processing subscriber nodes.
- Each Unified CM cluster can support configuration and registration for a maximum of 40,000 secured or unsecured endpoints. For additional information about Unified CM capacity planning, including per-platform sizing limits, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.
- When deploying a two-node cluster, Cisco recommends that you do not exceed 1,250 users in the cluster. Above 1,250 users, Cisco recommends a dedicated publisher and separate nodes for primary and backup call processing subscribers.
- Cisco recommends using the same VM configuration for all nodes in a cluster. Mixing Unified CM VM configurations is allowed, but there are design implications and limits. For more details, refer to the section on Mixing Unified CM VM Configurations, page 9-8.
- 2:1 redundancy is not supported when using the 10,000-user VM configuration option due to potential overload on the backup subscriber

- Use multiple physical ports in the hardware platform for the virtual machine network traffic, and use a minimum of two upstream switches to provide network connectivity redundancy. If using the VMware vSphere virtual switch, use VMware NIC teaming.
- Whenever possible, distribute the hardware platforms across multiple physical switches within the network and across multiple physical locations within the same network to minimize the impact of a switch failure or the loss of a particular network location.
- Deploy SRST or E-SRST on Cisco IOS routers at remote locations to provide fallback call processing services in the event that these locations lose connectivity to the Unified CM cluster.
- Cisco recommends that you leave voice activity detection (VAD) disabled within the Unified CM cluster. VAD is disabled by default in the Unified CM service parameters, and you should disable it on H.323 and SIP dial peers configured on Cisco IOS gateways by using the **no vad** command.
- Ensure that the Unified CM nodes are distributed across different physical servers so that backup or redundant subscriber nodes are on different physical servers than the primary subscriber nodes.
- Servers with slower CPUs can be restricted in terms of the Open Virtualization Archive (OVA) VM configuration that they support. We refer to those servers as servers with *restricted UC performance* CPUs. For example, the Cisco Unified CM 1000-user OVA VM configuration is the only Unified CM OVA VM configuration that can be installed on those servers. However, some smaller servers support only smaller VM configurations. For information on proper VM configuration selection as well as the use of the Cisco Collaboration Sizing Tool, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.
- Access to the USB and serial ports on the hardware platform is not supported with Unified CM virtual machines. Therefore, attaching fixed live audio sources for MoH, making a serial SMDI connection to a legacy voicemail system, or attaching a USB flash drive for writing log files are also not supported. The following alternative options are available:
  - For MoH live audio source feed, consider using Cisco IOS-based gateway multicast MoH for live audio source connectivity.
  - For saving system installation logs, use virtual floppy softmedia.
  - There is no support for SMDI serial connection.
- With Cisco Business Edition 6000, Unified CM is deployed as a single Unified CM publisher node that also handles call processing. To provide Unified CM redundancy, SRST can be deployed or additional hardware server(s) hosting Unified CM subscriber node(s) can be deployed.

Note More than two servers may be clustered for a BE6000 deployment to provide additional redundancy and/or geographic distribution; however, the capacity limits are not increased. For example, the total number of users across the cluster may not exceed 1,000 with BE6000M or BE6000H.

- If multiple Business Edition 6000 servers are required in the same deployment, distribute them across multiple physical switches.
- Use an uninterruptible power supply (UPS) to provide maximum availability, especially if the server has only one power supply.
- When deploying Business Edition 6000 with two servers for high availability, a Unified CM node should run on each server to provide high availability in case one of the servers fails. Furthermore, Cisco recommends configuring the Unified CM cluster with the subscriber node as the primary call processing server and the publisher node as the backup call processing server.

- With Cisco Business Edition 7000, Unified CM has the same rules, capacities, and design considerations as an enterprise (not part of Cisco Business Edition) Unified CM deployment.
- Applications that are not part of the Business Edition 6000 solution and that are running on separate hardware can be integrated to a Business Edition 6000 deployment, but you must ensure that those applications do not exceed the Business Edition 6000 capacity limits. For example, the overall BHCA and the number of contact center agents should not exceed the Business Edition 6000 capacity limit of Unified CM. For more information on the Business Edition 6000 capacity limits, refer to the chapter on Collaboration Solution Sizing Guidance, page 25-1. Also ensure that those applications support the Cisco Collaboration VM configuration provided by Business Edition 6000. For example, Cisco Unified Contact Center Enterprise requires the Unified CM 7.5k-user or larger VM configuration, so it cannot be integrated with a Unified CM deployment that is running on Business Edition 6000.

Cisco Unified CME

- Unified CME supports a maximum of 450 endpoints. However, depending on the Cisco IOS router model, endpoint capacity could be significantly lower. For additional information about Unified CME platforms and capacities, refer to the Cisco Unified Communications Manager Express compatibility information available at https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-express/products-device-support-tables-list.html.
- When possible, dual-attach the Unified CME router to the network using multiple IP interfaces to provide maximum network availability. Likewise, if multiple instances of Unified CME are required in the same deployment, distribute them across multiple physical switches or locations.
- When possible, deploy the Unified CME router with dual power supplies and/or an uninterruptible power supply (UPS) in order to provide maximum availability of the platform.

Computer Telephony Integration (CTI)

Cisco Computer Telephony Integration (CTI) extends the rich feature set available on Cisco Unified CM to third-party applications. The CTI-enabled applications improve user productivity, enhance the communication experience, and deliver superior customer service. At the desktop, Cisco CTI enables third-party applications to make calls from within Microsoft Outlook, open windows or start applications based on incoming caller ID, and remotely track calls and contacts for billing purposes. Cisco CTI-enabled server applications can intelligently route contacts through an enterprise network, provide automated caller services such as auto-attendant and interactive voice response (IVR), as well as capture media for contact recording and analysis.

CTI applications generally fall into one of two major categories:

- First-party applications — Monitor, control, and media termination

  First-party CTI applications are designed to register devices such as CTI ports and route points for call setup, tear-down, and media termination. Because these applications are directly in the media path, they can respond to media-layer events such as in-band DTMF. Interactive voice response and Cisco Attendant Console are examples of first-party CTI applications that monitor and control calls while also interacting with call media.

- Third-party application — Monitor and control

  Third-party CTI applications can also monitor and control calls, but they do not directly control media termination.

  - Monitoring applications

    A CTI application that monitors the state of a Cisco IP device is called a monitoring application. A busy-lamp-field application that displays on-hook/off-hook status or uses that information to indicate a user's availability in the form of Presence are both examples of third-party CTI monitoring applications.

  - Call control applications

    Any application that uses Cisco CTI to remotely control a Cisco IP device using out-of-band signaling is a call control application. Cisco Jabber, when configured to remotely control a Cisco IP device, is a good example of a call control application.

  - Monitor + call control applications

    These are any CTI applications that monitor and control a Cisco IP device. Cisco Unified Contact Center Enterprise is a good example of a combined monitor and control application because it monitors the status of agents and controls agent phones through the agent desktop.

Note While the distinction between a monitor, call control, and monitor + control application is called out here, this granularity is not exposed to the application developer. All CTI applications using Cisco CTI are enabled for both monitoring and control.

The following devices can be monitored or controlled through CTI:

- CTI Route Point
- CTI Port
- Cisco Unified IP Phones supporting CTI
- CTI Remote Device

CTI Remote Device provides the ability for a CTI application to have monitoring and limited call control capabilities over phones that do not support CTI, such as traditional PSTN phones, mobile phones, third-party phones, or phones attached to a third-party PBX.

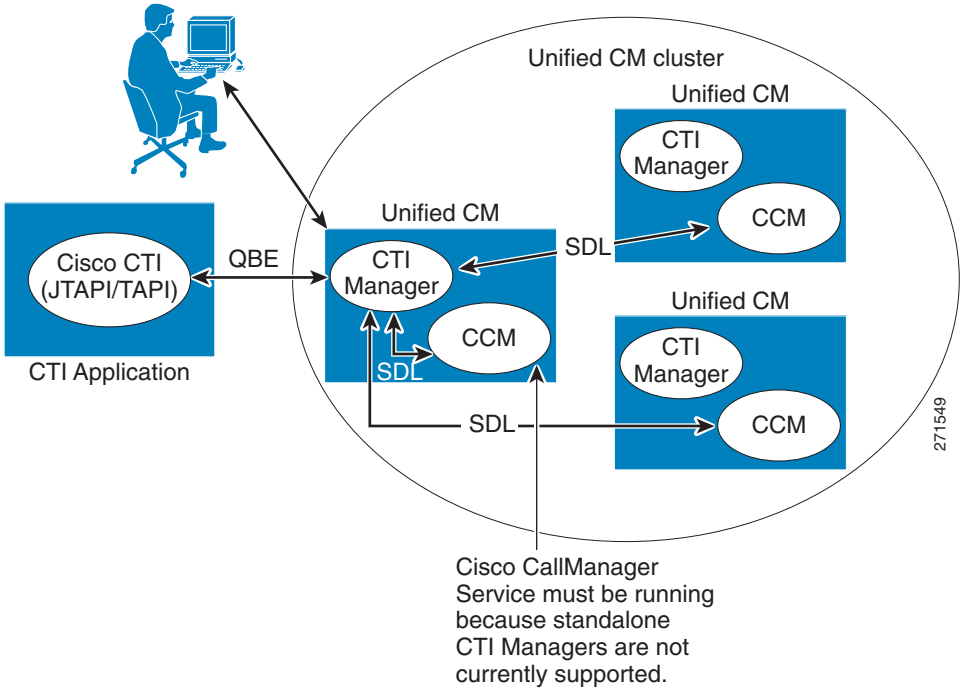
CTI Architecture

Cisco CTI consists of the following components (see Figure 9-9), which interact to enable applications to take advantage of the telephony feature set available in Cisco Unified CM:

- CTI-enabled application — Cisco or third-party application written to provide specific telephony features and/or functionality.
- JTAPI and TAPI — Two standard interfaces supported by Cisco CTI. Developers can choose to write applications using their preferred method library.
- Unified JTAPI and Unified TSP Client — Converts external messages to internal Quick Buffer Encoding (QBE) messages used by Cisco Unified CM.
- Quick Buffer Encoding (QBE) — Unified CM internal communication messages.
- Provider — A logical representation of a connection between the application and CTI Manager, used to facilitate communication. The provider sends device and call events to the application while accepting control instructions that allow the application to control the device remotely.

- Signaling Distribution Layer (SDL) — Unified CM internal communication messages.
- Publisher and subscriber — Cisco Unified Communications Manager (Unified CM) server nodes.
- CCM — The Cisco CallManager Service (ccm.exe), the telephony processing engine.
- CTI Manager (CTIM) — A service that runs on one or more Unified CM subscribers operating in primary/secondary mode and that authenticates and authorizes telephony applications to control and/or monitor Cisco IP devices.

Figure 9-9 Cisco CTI Architecture



Once an application is authenticated and authorized, the CTIM acts as the broker between the telephony application and the Cisco CallManager Service. (This service is the call control agent and should not be confused with the overall product name Cisco Unified Communications Manager.) The CTIM responds to requests from telephony applications and converts them to Signaling Distribution Layer (SDL) messages used internally in the Unified CM system. Messages from the Cisco CallManager Service are also received by the CTIM and directed to the appropriate telephony application for processing.

The CTIM may be activated on any of the Unified CM subscriber nodes in a cluster that have the Cisco CallManager Service active. This allows up to eight CTIMs to be active within a Unified CM cluster. Standalone CTIMs are currently not supported.

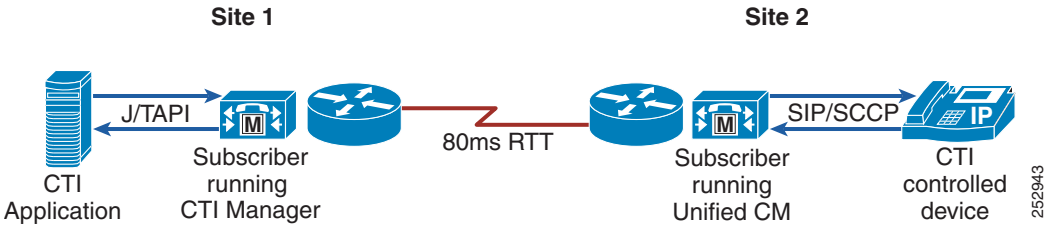
CTI Applications and Clustering Over the WAN

Deployments that employ clustering over the WAN are supported in the following two scenarios:

- CTI Manager over the WAN (see Figure 9-10)

  In this scenario, the CTI application and its associated CTI Manager are on one side of the WAN (Site 1), and the monitored or controlled devices are on the other side, registered to a Unified CM subscriber (Site 2). The round-trip time (RTT) must not exceed the currently supported limit of 80 ms for clustering over the WAN. To calculate the necessary bandwidth for CTI traffic, use the formula in the section on Local Failover Deployment Model, page 10-47. Note that this bandwidth is in addition to the Intra-Cluster Communication Signaling (ICCS) bandwidth calculated as described in the section on Local Failover Deployment Model, page 10-47, as well as any bandwidth required for audio (RTP traffic).

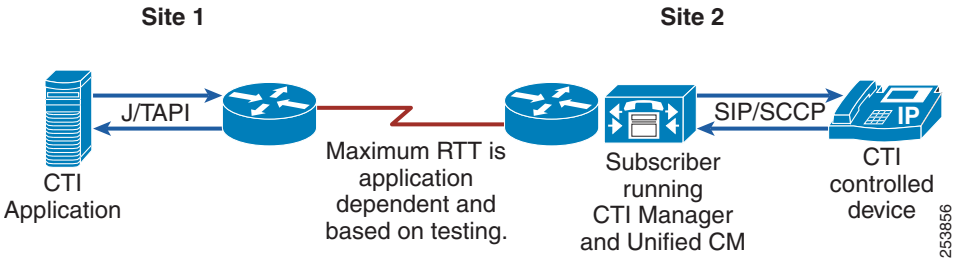
Figure 9-10 CTI Over the WAN



- TAPI and JTAPI applications over the WAN (CTI application over the WAN; see Figure 9-11)

  In this scenario, the CTI application is on one side of the WAN (Site 1), and its associated CTI Manager is on the other side (Site 2). In this scenario, it is up to the CTI application developer or provider to ascertain whether or not their application can accommodate the RTT as implemented. In some cases failover and failback times might be higher than if the application is co-located with its CTI Manager. In those cases, the application developer or provider should provide guidance as to the behavior of their application under these conditions.

Figure 9-11 JTAPI Over the WAN



Note Support for TAPI and JTAPI over the WAN is application dependent. Both customers and application developers or providers should ensure that their applications are compatible with any such deployment involving clustering over the WAN.

Capacity Planning for CTI

The maximum number of supported CTI-controlled devices is 40,000 per cluster. For more information on CTI capacity planning, including per-platform node and cluster CTI capacities as well as CTI resource calculation formulas and examples, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

High Availability for CTI

This section provides some guidelines for provisioning CTI for high availability.

CTI Manager

CTI Manager must be enabled on at least one and possibly all call processing subscribers within the Unified CM cluster. The client-side interfaces (TAPI TSP or JTAPI client) allow for two IP addresses each, which then point to Unified CM server nodes running the CTIM service. For CTI application redundancy, Cisco recommends having the CTIM service activated on at least two Unified CM server nodes in a cluster, as shown in Figure 9-12.

Redundancy, Failover, and Load Balancing

For CTI applications that require redundancy, the TAPI TSP or JTAPI client can be configured with two IP addresses, thereby allowing an alternate CTI Manager to be used in the event of a failure. It should be noted that this redundancy is not stateful in that no information is shared and/or made available between the two CTI Managers, and therefore the CTI application will have some degree of re-initialization to go through, depending on the exact nature of the failover.

When a CTI Manager fails-over, just the CTI application login process is repeated on the now-active CTI Manager. Whereas, if the Unified CM server node itself fails, then the re-initialization process is longer due to the re-registration of all the devices from the failed Unified CM to the now-active Unified CM, followed by the CTI application login process.

For CTI applications that require load balancing or that could benefit from this configuration, the CTI application can simply connect to two CTI Managers simultaneously, as shown in Figure 9-12.

Figure 9-12 Redundancy and Load Balancing

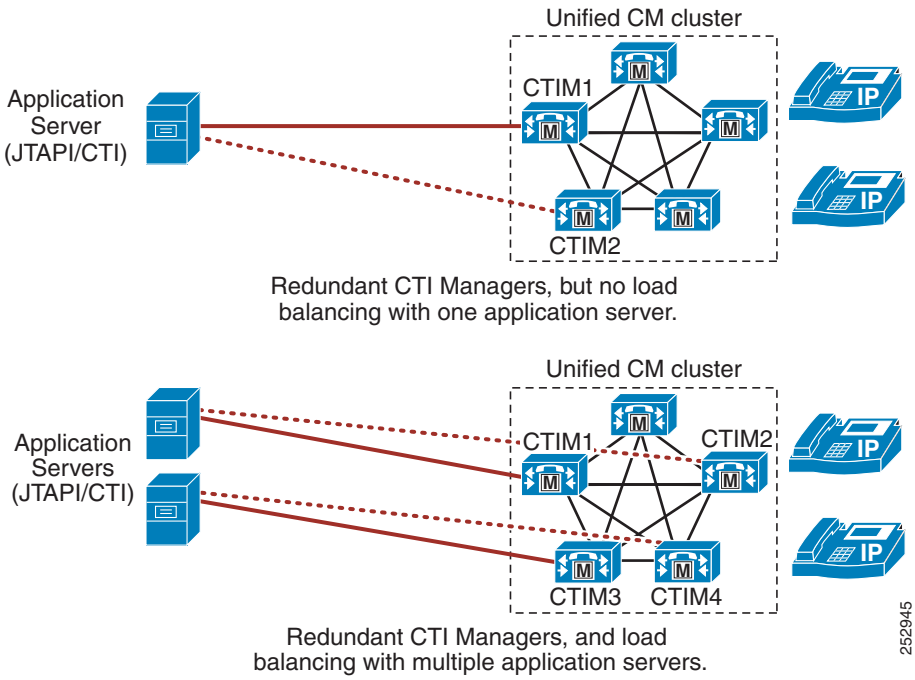


Figure 9-13 shows an example of this type of configuration for Cisco Unified Contact Center Enterprise (Unified CCE). This type of configuration has the following characteristics:

- Unified CCE uses two Peripheral Gateways (PGs) for redundancy.
- Each PG logs into a different CTI Manager.
- Only one PG is active at any one time.

Figure 9-13 CTI Redundancy with Cisco Unified Contact Center Enterprise

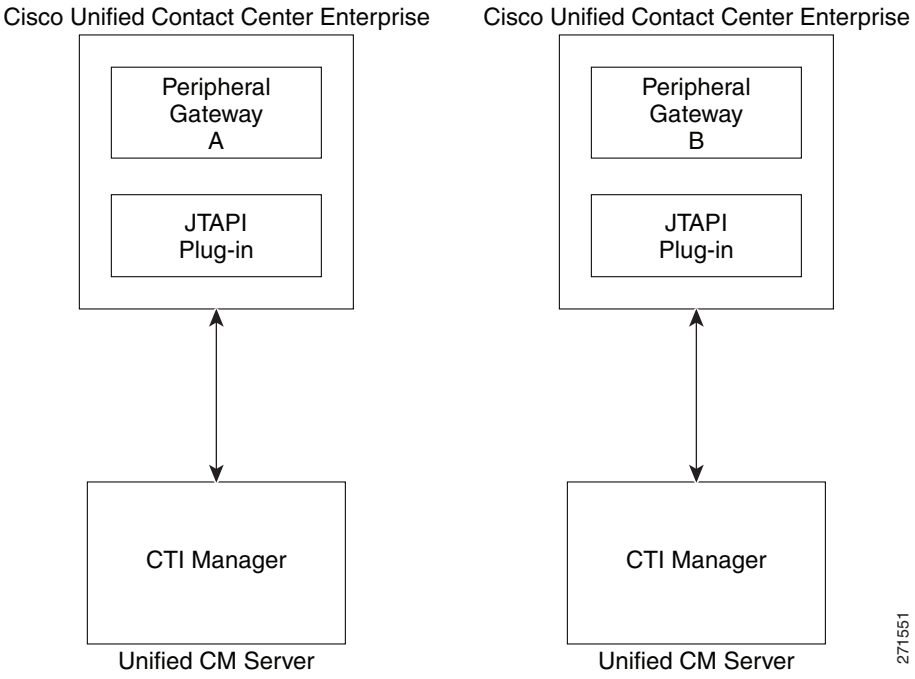
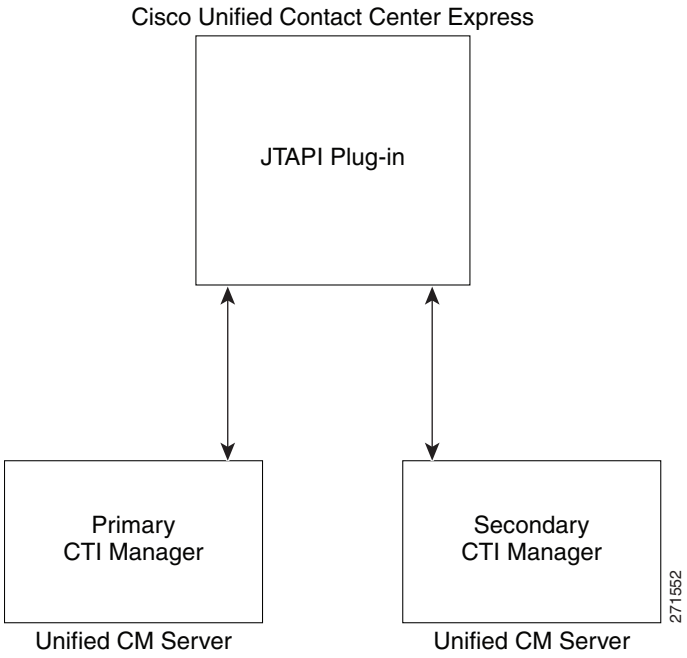


Figure 9-14 shows an example of this type of configuration for Cisco Unified Contact Center Express (Unified CCX). This type of configuration has the following characteristics:

- Unified CCX has two IP addresses configured, one for each CTI Manager.
- If connection to the primary CTI Manager is lost, Unified CCX fails-over to its secondary CTI Manager.

Figure 9-14 CTI Redundancy with Cisco Unified Contact Center Express



Implementation

For guidance and support on writing applications, application developers should consult the Cisco Developer Network (DevNet), located at

https://developer.cisco.com/site/devnet/home

Integration of Multiple Call Processing Agents

To integrate multiple Unified CM clusters together or to integrate Unified CM clusters with the Cisco TelePresence Video Communication Server (VCS), use Cisco Unified CM Session Management Edition (SME). SME is the recommended trunk and dial plan aggregation platform in multi-site distributed call processing deployments. SME is essentially a Unified CM cluster with trunk interfaces only and no IP endpoints. It enables aggregation of multiple Unified Communications systems, referred to as *leaf* systems.

Unified CM Session Management Edition may also be used to connect to third-party unified communications systems such as PSTN connections, PBXs, and centralized unified communications applications.

For more information on SME, see the section on Unified CM Session Management Edition, page 10-26.

Direct integration of multiple call processing agents is also possible. This section explains the requirements for interoperability and internetworking of Cisco Unified CM with Cisco Unified Communications Manager Express (Unified CME) using SIP trunking protocol in a multisite IP telephony deployment. This section highlights the recommended deployments between phones controlled by Unified CM and phones controlled by Unified CME.

This section covers the following topics:

- Overview of Interoperability Between Unified CM and Unified CME, page 9-36
- Unified CM and Unified CME Interoperability via SIP in a Multisite Deployment with Distributed Call Processing, page 9-38

Cisco Unified CM and Cisco Unified Communications Manager Express (Unified CME) could also be integrated using H.323, but this section does not cover this integration in detail. For more information on the H.323 integration, refer to the *Cisco Collaboration 9.x SRND*, available at

https://www.cisco.com/go/srnd

Overview of Interoperability Between Unified CM and Unified CME

Either H.323 or SIP can be used as a trunking protocol to interconnect Unified CM and Unified CME. When deploying Unified CM at the headquarters or central site in conjunction with one or more Unified CME systems for branch offices, network administrators must choose either the SIP or H.323 protocol after careful consideration of protocol specifics and supported features across the WAN trunk. Using H.323 trunks to connect Unified CM and Unified CME has been the predominant method in past years, until more enhanced capabilities for SIP phones and SIP trunks were added in Unified CM and Unified CME. This section first describes some of the features and capabilities that are independent of the trunking protocol for Unified CM and Unified CME interoperability, then it explains some of the most common design scenarios and best practices for using SIP trunks.

Call Types and Call Flows

In general, Unified CM and Unified CME interworking allows all combination of calls from SCCP IP phones to SIP IP phones, or vice versa, across a SIP trunk or H.323 trunk. Calls can be transferred (blind or consultative) or forwarded back and forth between the Unified CM and Unified CME SIP and/or SCCP IP phones.

When connected to Unified CM via H.323 trunks, Unified CME can auto-detect Unified CM calls. When a call terminating on Unified CME is transferred or forwarded, Unified CME regenerates the call and routes the call appropriately to another Unified CME or Unified CM by hairpinning the call.

Unified CME hairpins the call legs from Unified CM for the VoIP calls across SIP or H.323 trunks when needed. For more information on allowing auto-detection on a non-H.450 supported Unified CM network and for enabling or disabling supplementary services for H450.2, H450.3, or SIP, refer to the Unified CME product documentation available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-express/tsd-products-support-series-home.html

When connected to Unified CM via SIP trunks, Unified CME does not auto-detect Unified CM calls. By default, Unified CME always tries to redirect calls using either a SIP Refer message for call transfer or a SIP 302 Moved Temporarily message for call forward; if that fails, Unified CME will then try to hairpin the call.

Music on Hold

While Unified CM can be enabled to stream MoH in both G.711 and G.729 formats, Unified CME streams MoH only in G.711 format. Therefore, when Unified CME controls the MoH audio on a call placed on hold, it requires a transcoder to transcode between a G.711 MoH stream and a G.729 call leg.

Instant and Permanent Hardware Conferencing

Hardware DSP resources are required for both instant and permanent conferences. Whether connected via SIP, H.323, or PSTN, both Unified CM and Unified CME phones can be invited or added to an instant conference to become conference participants as along as the phones are reachable from the network. When calls are put on hold during an active conference session, music will not be heard by the conference participants in the conference session.

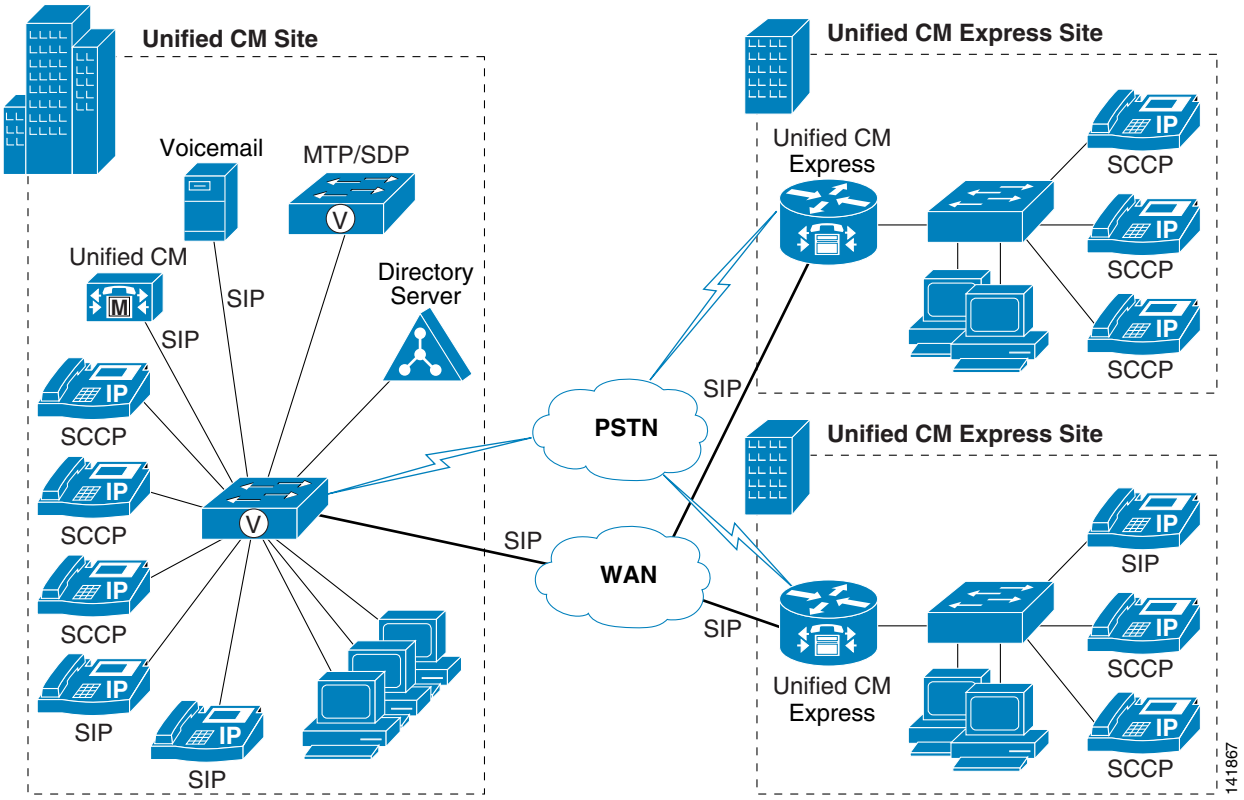
For information on required and supported DSP resources and the maximum number of conference participants allowed for instant or permanent conferences, refer to the Unified CME product documentation available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-express/tsd-products-support-series-home.html

Unified CM and Unified CME Interoperability via SIP in a Multisite Deployment with Distributed Call Processing

Unified CM can communicate directly with Unified CME using a SIP interface. Figure 9-15 shows a Cisco Unified Communications multisite deployment with Unified CM networked directly with Cisco Unified CME using a SIP trunk.

Figure 9-15 Multisite Deployment with Unified CM and Unified CME Using SIP Trunks



Best Practices

Follow these guidelines and best practices when using the deployment model illustrated in Figure 9-15:

- Configure a SIP Trunk Security Profile with **Accept Replaces Header** selected.
- Configure a SIP trunk on Unified CM using the SIP Trunk Security Profile created, and also specify a ReRouting CSS. The ReRouting CSS is used to determine where a SIP user (transferor) can refer another user (transferee) to a third user (transfer target) and which features a SIP user can invoke using the SIP 302 Redirection Response and INVITE with Replaces.
- For SIP trunks there is no need to enable the use of media termination points (MTPs) when using SCCP endpoints on Unified CME. However, SIP endpoints on Unified CME require the use of media termination points on Unified CM to be able to handle delayed offer/answer exchanges with the SIP protocol (that is, the reception of INVITEs with no Session Description Protocol).

- Route calls to Unified CME via a SIP trunk using the Unified CM dial plan configuration (route patterns, route lists, and route groups).
- Use Unified CM device pools and regions to configure a G.711 codec within the site and the G.729 codec for remote Unified CME sites.
- Configure the **allow-connections sip to sip** command under **voice services voip** on Unified CME to allow SIP-to-SIP call connections.
- For SIP endpoints, configure the **mode cme** command under **voice register global**, and configure **dtmf-relay rtp-nte** under the **voice register pool** commands for each SIP phone on Unified CME.
- For SCCP endpoints, configure the **transfer-system full-consult** command and the **transfer-pattern .T** command under **telephony-service** on Unified CME.
- Configure the SIP WAN interface voip dial-peers to forward or redirect calls, destined for Unified CM, with **session protocol sipv2** and **dtmf-relay** [**sip-notify** | **rtp-nte**] on Unified CME.

Design Considerations

This section first covers some characteristics and design considerations for Unified CM and Unified CME interoperability via SIP in some main areas such as supplementary services for call transfer and forward, presence service for busy lamp field (BLF) notification for speed-dial buttons and directory call lists, and out-of-dialog (OOD-Refer) for integration with partner applications and third-party phone control for click-to-dial between the Unified CM phones and Unified CME phones. The section also covers some general design considerations for Unified CM and Unified CME interoperability via SIP.

Supplementary Services

SIP Refer or SIP 302 Moved Temporarily messages can be used for supplementary services such as call transfer or call forward on Unified CME or Unified CM to instruct the transferee (referee) or phone being forwarded (forwardee) to initiate a new call to the transfer-to (refer-to) target or forward-to target. No hairpinning is needed for call transfer or call forward scenarios when the SIP Refer or SIP 302 Moved Temporarily message is supported.

However, **supplementary-service** must be disabled if there are certain extensions that have no DID mapping or if Unified CM or Unified CME does not have a dial plan to route the call to the DID in the SIP 302 Moved Temporarily message. When **supplementary-service** is disabled, Unified CME hairpins the calls or sends a re-invite SIP message to Unified CM to replace the media path to the new called party ID. Both signaling and media are hairpinned, even when multiple Unified CMEs are involved for further call forwards. The **supplementary-service** can also be disabled for transferred calls. In this case, the SIP Refer message will not be sent to Unified CM, but the transferee (referee) party and transfer-to party (refer-to target) are hairpinned.

Note Supplementary services can be disabled with the command **no supplementary-service sip moved-temporarily** or **no supplementary-service sip refer** under **voice service voip** or **dial-peer voice xxxx voip**.

The following examples illustrate the call flows when supplementary services are disabled:

- Unified CM phone B calls Unified CME phone A, which is set to call-forward (all, busy, or no answer) to phone C (either a Unified CM phone, a Unified CME phone on the same or different Unified CME, or a PSTN phone).

  Unified CME does not send the SIP 302 Moved Temporarily message to Unified CM, but hairpins the call between Unified CM phone B and phone C.

- Unified CM phone B calls Unified CME phone A, which transfer the call to phone C (either a Unified CM phone, a Unified CME phone, or a PSTN phone).

  Unified CME does not send the SIP Refer message to Unified CM, but hairpins the call between Unified CM phone B and phone C.

General Design Considerations for Unified CM and Unified CME Interoperability via SIP

- Disable **supplementary-service** if SIP 302 Moved Temporarily or SIP Refer messages are not supported by Unified CM, otherwise Unified CM cannot route the call to the transfer-to or forward-to target.
- In a SIP-to-SIP call scenario, a Refer message is sent by default from the transferor to the transferee, the transferee sets up a new call to the transfer-to target, and the transferor hears ringback tone by default while waiting for the transfer at connect. If **supplementary-service** is disabled on Unified CME, Unified CME will provide in-band ringback tone right after the call between the transferee and transfer-to target is connected.
- Presence service is supported on Unified CM and Unified CME via SIP trunk only.
- The OOD-Refer feature allows third-party applications to connect two endpoints on Unified CM or Unified CME through the use of the SIP REFER method. Consider the following factors when using OOD-Refer:
  - Both Unified CM and Unified CME must be configured to enable the OOD-Refer feature.
  - Call Hold, Transfer, and Conference are not supported during an OOD-Refer transaction, but they are not blocked by Unified CME.
  - Call transfer is supported only after the OOD-Refer call is in the connected state and not before the call is connected; therefore, call transfer-at-alert is not supported.
- Control signaling in TLS is supported, but SRTP is not supported over the SIP trunk.
- SRTP over a SIP trunk is a gateway feature in Cisco IOS for Unified CM. SRTP support is not available with Unified CM and Unified CME interworking via SIP trunks.

Note When multiple PSTN connections exist (one for Unified CM and one for Unified CME), fully attended transfer between a Unified CM endpoint and a Unified CME endpoint to a PSTN endpoint will fail. The recommendation is to use blind transfer when using multiple PSTN connections, and it is configured under **telephony-service** as **transfer-system full-blind**.

Collaboration Deployment Models

Revised: March 1, 2018

This chapter describes the deployment models for Cisco Collaboration Systems.

Earlier versions of this chapter based the deployment models discussion exclusively on the call processing deployment models for Cisco Unified Communications Manager (Unified CM). The current version of this chapter offers design guidance for the entire Cisco Unified Communications and Collaboration System, which includes much more than just the call processing service.

For design guidance with earlier releases of Cisco Unified Communications, refer to the Cisco Unified Communications Solution Reference Network Design (SRND) documentation available at

https://www.cisco.com/go/srnd

What's New in This Chapter

Table 10-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 10-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Minor updates and corrections	Various sections of this chapter	March 1, 2018

Deploying Unified Communications and Collaboration

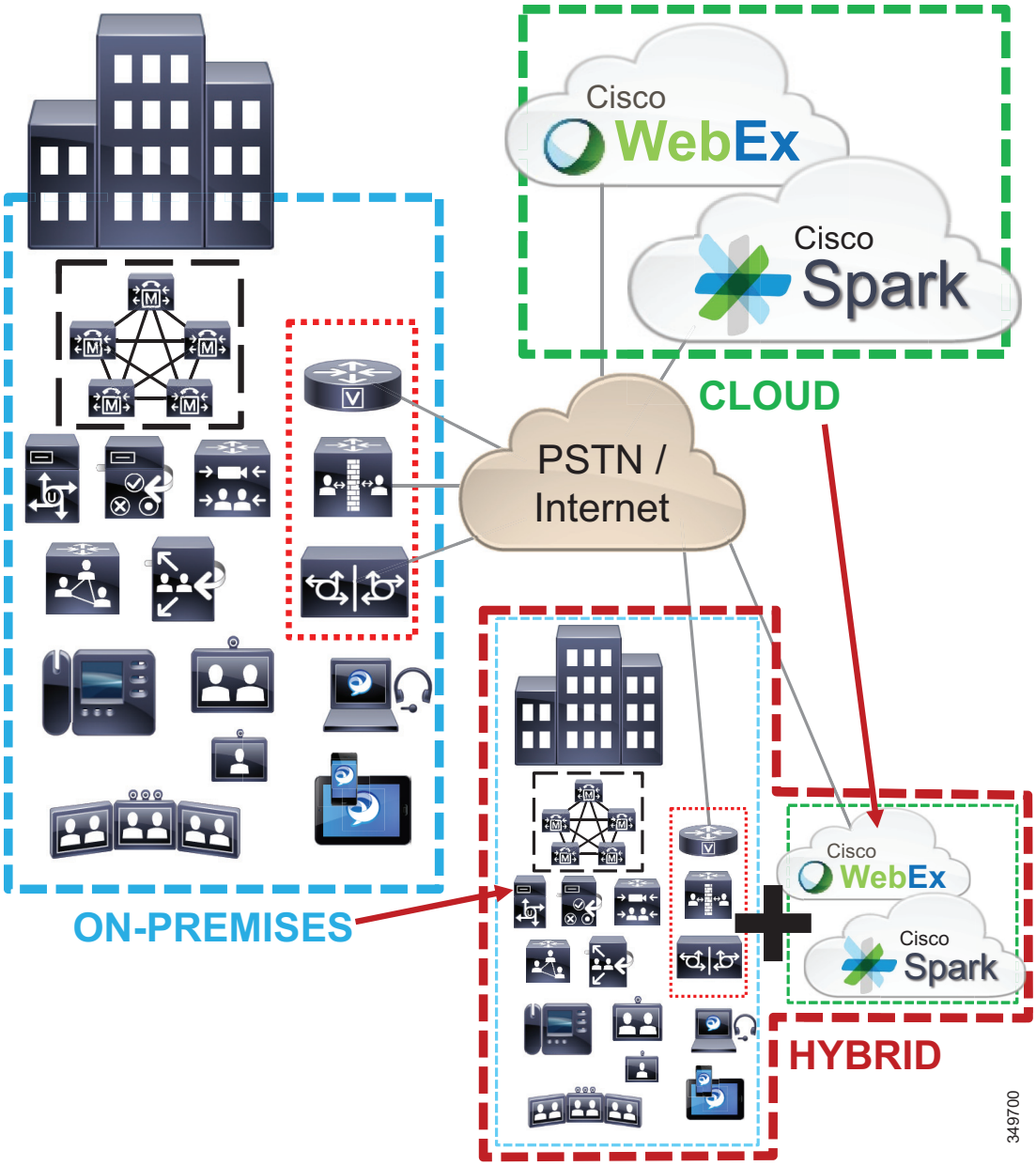
From its beginnings with Voice over IP (VoIP) and IP Telephony more than 15 years ago, and continuing today with Unified Communications and Collaboration, users expect to be able to meet and communicate in a variety ways using a range of devices with differing capabilities. Today, a Unified Communications and Collaboration system could start with the deployment of Jabber Clients for IM and Presence only and incrementally add voice, video, web conferencing, mobile voice applications, social media, video conferencing, and telepresence as required. A tightly integrated Unified Communications architecture is required as the number of devices and forms of communication available to a single Unified Communications user increases. Cisco's Unified Communications and Collaboration architecture has the flexibility and scale to meet the demands of a rapidly changing and expanding Unified Communications environment that will become more URI-centric as users with multiple Unified Communications devices wish to be identified by a single user name irrespective of the form of communication.

Enterprise Collaboration Deployments

Collaboration and Unified Communications (UC) deployments for enterprises have traditionally involved delivering applications and services from within the enterprise network boundary, or on-premises. Increasingly, collaboration and UC services are being delivered from the cloud using the "as a Service" ("aaS") paradigm (for example, Collaboration as a Service, Communication Platform as a Service, UC as a Service, and so forth). In this case, one or many services may be delivered from the cloud. In cases where enterprises desire the benefits of both on-premises services (such as existing investment, high quality voice and video calling, and so forth) and cloud services (such as continuous delivery or mobile and web delivery), those enterprises are most often implementing hybrid deployments with a combination of both on-premises and cloud-based collaboration applications and services.

As shown in Figure 10-1, collaboration applications and services may be delivered solely on-premises, solely in the cloud, or more and more commonly in combination as a set of hybrid service deployments.

Figure 10-1 Enterprise Collaboration Deployments: On-Premises, Cloud, and Hybrid



For example, with an on-premises deployment, collaboration applications are deployed within the enterprise premises to provide voice and video calling; text, voice, and video messaging; presence; and video conferencing and desk, screen, and content sharing. The applications and services include:

- Unified CM
- Unified CM IM and Presence
- Unity Connection
- Cisco Meeting Server and TelePresence Management Suite (TMS)

In the case of cloud deployments, collaboration services delivered from the cloud include voice and video calling, messaging, and meetings with video as well as content and screen sharing. Services delivered from the cloud include:

- Cisco Spark message, meeting, and call — Cisco Spark 1:1 and team messaging; audio, video, and web meetings; and voice and video calling – across mobile, web, and desktop platforms
- WebEx meet and messaging — WebEx Messenger and WebEx Meetings (mobile, web, and desktop; collaboration meeting room (CMR); and so forth)

For more information on Cisco Webex, refer to the documentation at https://help.webex.com/welcome. For more information on Cisco Spark refer to the documentation at https://support.ciscospark.com/.

Additional cloud implementations of collaboration can include collaboration platform-based services as provided by third-party managed service providers and integrators that deliver traditional on-premises collaboration applications and services from the cloud. Cisco Hosted Collaboration Solution (HCS) is an example of this type of cloud platform-based services.

For more information on Cisco Hosted Collaboration Solution, refer to the documentation available at https://www.cisco.com/c/en/us/solutions/hosted-collaboration-solution/index.html.

Also, implementations can include both on-premises and cloud-based services to provide hybrid deployments that enable organizations to leverage the advantages of both delivery mechanisms. Examples of hybrid deployments include:

- Enterprise calling with cloud messaging and conferencing — Cisco Unified CM and Unity Connection on-premises for call control and voice and video messaging, with WebEx Messenger and WebEx Meetings for IM and presence, voice and video, as well as web-based conferencing, including permanent collaboration meeting rooms.
- Enterprise calling with cloud calling and messaging — Cisco Unified CM on-premises for voice and video calling, with Cisco Spark messaging and meet as well as hybrid services for voice and video calling integration between on-premises and the cloud.

The deployment models in this chapter predominately cover on-premises deployments. However, in all cases, cloud-based applications and services can be integrated with the various deployment models to enable hybrid deployments.

Deployment Model Architecture

In general terms, the deployment model architecture follows that of the enterprise it is deployed to serve. Deployment models describe the reference architecture required to satisfy the Unified Communications needs of well-defined, typical topologies of enterprises. For example, a centralized call processing deployment model caters to enterprises whose operational footprint is based on multiple sites linked to one or few centralized headquarters offices.

In some cases, the deployment model of a technology will depart from that of the enterprise, due to technological constraints. For example, if an enterprise has a single campus whose scale exceeds that of a single service instance, such as a call processing service provided by Cisco Unified Communications Manager, then a single campus might require more than a single instance of a call processing cluster or a single messaging product.

Another option for customers who exceed the sizing limits of a standard cluster is to consider deploying a megacluster, which can provide increased scalability. For more information about megaclusters, see Megacluster, page 9-25.

Note Unless otherwise specified, all information contained within this SRND that relates to call processing deployments (including capacity, high availability, and general design considerations) applies only to a standard cluster with up to eight call processing subscriber nodes.

Summary of Unified Communications Deployment Models

This chapter discusses three basic on-premises deployment models for Unified Communications and Collaboration:

- Campus deployment model

  Where the Unified Communications and Collaboration services, their associated endpoints, gateways, border controllers, media resources, and other components are all located on a single high speed LAN or MAN.

- Centralized deployment model

  Where the Unified Communications and Collaboration services are located in a central campus site or data center, but the endpoints, gateways, media resources, and other components are distributed across multiple remote sites that are interconnected by a QoS-enabled WAN.

- Distributed deployment model

  Where multiple campus and/or centralized deployments are interconnected by means of a trunk and dial plan aggregation platform, such as a Cisco Unified Communications Manager Session Management Edition cluster, over a QoS-enabled WAN.

There are an infinite number of variations on these three basic deployment models, such as deployments with centralized or distributed PSTN access and services, but the basic design guidance provided in this chapter still applies to the majority of them.

High Availability for Deployment Models

Unified Communications services offer many capabilities aimed at achieving high availability. They may be implemented in various ways, such as:

- Failover redundancy

  For services that are considered essential, redundant elements should be deployed so that no single point of failure is present in the design. The redundancy between the two (or more) elements is automated. For example, the clustering technology used in Cisco Unified Communications Manager (Unified CM) allows for up to three servers to provide backup for each other. This type of redundancy may cross technological boundaries. For example, a phone may have as its first three preferred call control agents, three separate Unified CM servers belonging to the same call processing cluster. As a fourth choice, the phone can also be configured to rely on a Cisco IOS router for call processing services.

- Redundant links

  In some instances, it is advantageous to deploy redundant IP links, such as IP WAN links, to guard against the failure of a single WAN link.

- Geographical diversity

  Some products support the distribution of redundant service nodes across WAN links so that, if an entire site is off-line (such as would be the case during an extended power outage exceeding the capabilities of provisioned UPS and generator backup systems), another site in a different location can ensure business continuance.

Capacity Planning for Deployment Models

The capacities of various deployment models are typically integrally linked to the capacities of the products upon which they are based. Where appropriate in this chapter, capacities are called out. For some of the products supporting services covered in more detail in other sections of this document, the capacities of those products are discussed in their respective sections.

Common Design Criteria

Across all technologies that make up the Cisco Unified Communications System, the following common set of criteria emerges as the main drivers of design:

Size

In this context, size generally refers to the number of users, which translates into a quantity of IP telephones, voice mail boxes, presence watchers, and so forth. Size also can be considered in terms of processing capacity for sites where few (or no) users are present, such as data centers.

Network Connectivity

The site's connectivity into the rest of the system has three main components driving the design:

- Bandwidth enabled for Quality of Service (QoS)
- Latency
- Reliability

These components are often considered adequate in the Local Area Network (LAN): QoS is achievable with all LAN equipment, bandwidth is typically in the Gigabit range, latency is minimal (in the order of a few milliseconds), and excellent reliability is the norm.

The Metropolitan Area Network (MAN) often approaches the LAN in all three dimensions: bandwidth is still typically in the multiple Megabit range, latency is typically in the low tens of milliseconds, and excellent reliability is common. Packet treatment policies are generally available from MAN providers, so that end-to-end QoS is achievable.

The Wide Area Network (WAN) generally requires extra attention to these components: the bandwidth is at a cost premium, the latencies may depend not only on effective serialization speeds but also on actual transmission delays related to physical distance, and the reliability can be impacted by a multitude of factors. The QoS performance can also require extra operational costs and configuration effort.

Bandwidth has great influence on the types of Unified Communications services available at a site, and on the way these services are provided. For example, if a site serving 20 users is connected with 1.5 Mbps of bandwidth to the rest of the system, the site's voice, presence, instant messaging, email, and video services can readily be hosted at a remote datacenter site. If that same site is hosting 1000 users, some of the services would best be hosted locally to avoid saturating the comparatively limited bandwidth with signaling and media flows. Another alternative is to consider increasing the bandwidth to allow services to be delivered across the WAN from a remote datacenter site.

The influence of latency on design varies, based on the type of Unified Communications service considered for remote deployment. If a voice service is hosted across a WAN where the one-way latency is 200 ms, for example, users might experience issues such as delay-to-dialtone or increased media cut-through delays. For other services such as presence, there might be no problem with a 200 ms latency.

Reliability of the site's connectivity into the rest of the network is a fundamental consideration in determining the appropriate deployment model for any technology. When reliability is high, most Unified Communications components allow for the deployment of services hosted from a remote site; when reliability is inconsistent, some Unified Communications components might not perform reliably when hosted remotely; if the reliability is poor, co-location of the Unified Communications services at the site might be required.

High Availability Requirements

The high availability of services is always a design goal. Pragmatic design decisions are required when balancing the need for reliability and the cost of achieving it. The following elements all affect a design's ability to deliver high availability:

- Bandwidth reliability, directly affecting the deployment model for any Unified Communications service
- Power availability

  Power loss is a very disruptive event in any system, not only because it prevents the consumption of services while the power is out, but also because of the ripple effects caused by power restoration. A site with highly available power (for example, a site whose power grid connection is stable, backed-up by uninterruptible power supplies (UPSs) and by generator power) can typically be chosen to host any Unified Communications service. If a site has inconsistent power availability, it would not be judicious to use it as a hosting site.
- Environmental factors such as heat, humidity, vibration, and so forth

  Some Unified Communications services are delivered though the use of equipment such as servers that require periodical maintenance. Some Unified Communications functions such as the hosting of Unified Communications call agent servers are best deployed at sites staffed with qualified personnel.

Site-Based Design Guidance

Throughout this document, design guidance is organized along the lines of the various Unified Communications services and technologies. For instance, the call processing chapter contains not only the actual description of the call processing services, but also design guidance pertaining to deploying IP phones and Cisco Unified Communications servers based on a site's size, network connectivity, and high availability requirements. Likewise, the call admission control chapter focuses on the technical explanation of that technology while also incorporating site-based design considerations.

Generally speaking, most aspects of any given Unified Communications service or technology are applicable to all deployments, no matter the site's size or network connectivity. When applicable, site-based design considerations are called out. Services can be centralized, distributed, inter-networked, and geographically diversified.

Centralized Services

For applications where enterprise branch sites are geographically dispersed and interconnected over a Wide Area Network, the Cisco Unified Communications services can be deployed at a central location while serving endpoints over the WAN connections. For example, the call processing service can be deployed in a centralized manner, requiring only IP connectivity with the remote sites to deliver telephony services. Likewise, voice messaging services, such as those provided by the Cisco Unity Connection platform, can also be provisioned centrally to deliver services to endpoints remotely connected across an IP WAN.

Centrally provisioned Unified Communications services can be impacted by WAN connectivity interruptions; for each service, the available local survivability options should be planned. As an example, the call processing service as offered by Cisco Unified CM can be configured with local survivability functionality such as Survivable Remote Site Telephony (SRST) or Enhanced SRST. Likewise, a centralized voice messaging service such as that of Cisco Unity Connection can be provisioned to allow remote sites operating under SRST to access local voicemail services using Unity Connection Survivable Remote Site Voicemail (SRSV).

The centralization of services need not be uniform across all Unified Communications services. For example, a system can be deployed where multiple sites rely on a centralized call processing service, but can also be provisioned with a de-centralized (distributed) voice messaging service such as Cisco Unity Express. Likewise, a Unified Communications system could be deployed where call processing is provisioned locally at each site through Cisco Unified Communications Manager Express (Unified CME), with a centralized voice messaging service such as Cisco Unity Connection.

In many cases, the main criteria driving the design for each service are the availability and quality of the IP network between sites. The centralization of Unified Communications services offers advantages of economy of scale in both capital and operational expenses associated with the hosting and operation of equipment in situations where the IP connectivity between sites offers the following characteristics:

- Enough bandwidth for the anticipated traffic load, including peak hour access loads such as those generated by access to voicemail, access to centralized PSTN connectivity, and inter-site on-net communications including voice and video
- High availability, where the WAN service provider adheres to a Service Level Agreement to maintain and restore connectivity promptly
- Low latency, where local events at the remote site will not suffer if the round-trip time to the main central site imparts some delays to the system's response times

Also, when a given service is deployed centrally to serve endpoints at multiple sites, there are often advantages of feature transparency afforded by the use of the same processing resources for users at multiple sites. For example, when two sites are served by the same centralized Cisco Unified Communications Manager (Unified CM) cluster, the users can share line appearances between the two sites. This benefit would not be available if each site were served by different (distributed) call processing systems.

These advantages of feature transparency and economies of scale should be evaluated against the relative cost of establishing and operating a WAN network configured to accommodate the demands of Unified Communications traffic.

Distributed Services

Unified Communications services can also be deployed independently over multiple sites, in a distributed fashion. For example, two sites (or more) can be provisioned with independent call processing Cisco Unified CME nodes, with no reliance on the WAN for availability of service to their co-located endpoints. Likewise, sites can be provisioned with independent voice messaging systems such as Cisco Unity Express.

The main advantage of distributing Unified Communications services lies in the independence of the deployment approach from the relative availability and cost of WAN connectivity. For example, if a company operates a site in a remote location where WAN connectivity is not available, is very expensive, or is not reliable, then provisioning an independent call processing node such as Cisco Unified CME within the remote site will avoid any call processing interruptions if the WAN goes down.

Inter-Networking of Services

If two sites are provisioned with independent services, they can still be interconnected to achieve some degree of inter-site feature transparency. For example, a distributed call processing service provisioned through Cisco Unified CME can be inter-networked through SIP or H.323 trunks to permit IP calls between the sites. Likewise, separate instances of Cisco Unity Connection or Cisco Unity Express can partake in the same messaging network to achieve the routing of messages and the exchange of subscriber and directory information within a unified messaging network.

Geographical Diversity of Unified Communications Services

Some services can be provisioned in multiple redundant nodes across the IP WAN. Depending on the design and features in use, this can provide the possibility for continued service during site disruptions such as loss of power, network outages, or even compromises in the physical integrity of a site by catastrophic events such as a fire or earthquake.

To achieve such geographical diversity, the individual service must support redundant nodes as well as the deployment of these nodes across the latency and bandwidth constraints of the IP WAN. For example, the call processing service of Unified CM does support the deployment of a single cluster's call processing nodes across an IP WAN as long as the total end-to-end round-trip time between the nodes does not exceed 80 ms and an appropriate quantity of QoS-enabled bandwidth is provisioned. By contrast, Unified CME does not offer redundancy, and thus cannot be deployed in a geographically diverse configuration.

Table 10-2 summarizes the ability of each Cisco Unified Communications service to be deployed in the manners outlined above.

Table 10-2 Available Deployment Options for Cisco Unified Communications Services

Service	Centralized	Distributed	Inter-Networked	Geographical Diversity
Cisco Unified CM: • Enterprise Edition • Business Edition 6000 • Business Edition 7000	Yes	Yes	Yes	Yes
Cisco Business Edition 4000	Yes	No	No	No
Cisco Unified CME	No	Yes	Yes	No
Cisco Unity Express	No	Yes	Yes, through Voice Profile for Internet Mail (VPIM) networking	No
Cisco Unity Connection	Yes	Yes (One Cisco Unity Connection per site)	Yes, through VPIM networking	Yes
Cisco Emergency Responder	Yes	Yes (One Emergency Responder group per site)	Yes, through Emergency Responder clustering	Yes
Cisco IM and Presence	Yes	Yes (one Cisco IM and Presence Service per site)	Yes, through inter-domain federation	Yes
Cisco Unified Mobility	Yes	Yes, as Unified CM Single Number Reach	No	Yes
Cisco Expressway	Yes	Yes	Yes	Yes
Cisco Meeting Server	Yes	Yes	Yes	Yes

Because call processing is a fundamental service, the basic call processing deployment models are introduced in this chapter. For a detailed technical discussion on Cisco Unified Communications Manager call processing, refer to the chapter on Call Processing, page 9-1.

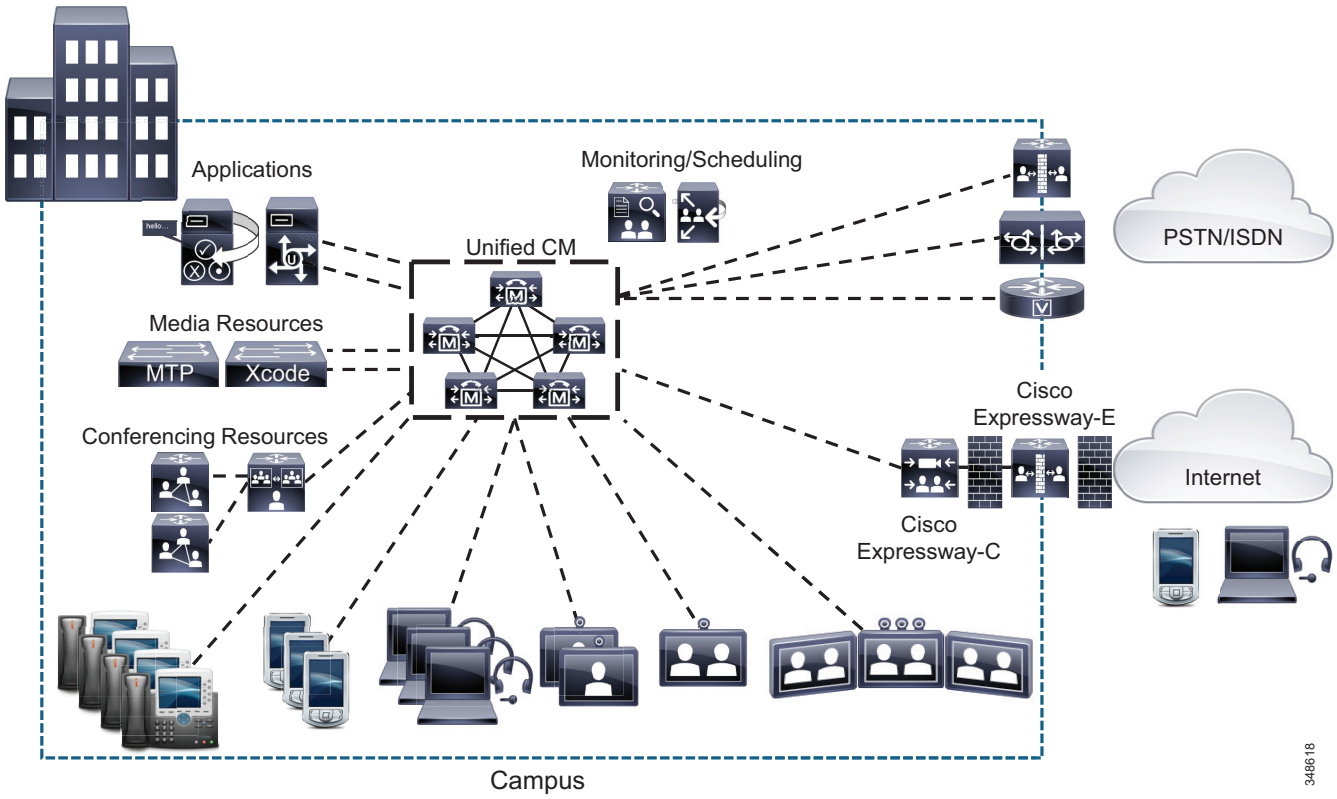
Design Characteristics and Best Practices for Deployment Models

This section describes the fundamental deployment models for Cisco Collaboration and Unified Communications systems, and it lists best practices for each model.

Campus Deployments

In this call processing deployment model, the Unified Communications services and the endpoints are co-located in the campus, and the QoS-enabled network between the service nodes, the endpoints, and applications is considered highly available, offering bandwidth in the gigabit range with less than 15 ms of latency end-to-end. Likewise, the quality and availability of power are very high, and services are hosted in an appropriate data center environment. Communications between the endpoints traverses a LAN or a MAN, and communications outside the enterprise goes over an external network such as the PSTN. An enterprise would typically deploy the campus model over a single building or over a group of buildings connected by a LAN or MAN. (See Figure 10-2.)

Figure 10-2 Example of a Campus Deployment



The campus model typically has the following design characteristics:

- Single Cisco Unified CM cluster (Enterprise or Business Edition 7000). Some campus call processing deployments may require more than one Unified CM cluster, for instance, if scale calls for more endpoints than can be serviced by a single cluster or if a cluster needs to be dedicated to an application such as a call center.
- Alternatively for smaller deployments, Cisco Business Edition 4000 or Business Edition 6000 may be deployed in the campus.
- Maximum of 40,000 configured and registered Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) IP phones, softphones, analog ports, video endpoints, SIP-based TelePresence endpoints and room-based TelePresence conferencing systems, mobile clients, and Cisco Virtualization Experience Clients (VXC) per Unified CM cluster.
- Maximum of 2,100 gateways and trunks (that is, the total number of H.323 gateways, H.323 trunks, digital MGCP devices, and SIP trunks) per Unified CM cluster.
- Trunks and/or gateways (IP or PSTN) for all calls to destinations outside the campus.
- Multipoint conferencing resources [multipoint control unit (MCU), TelePresence Server, or other multipoint resources] are required for multipoint conferencing.
- Co-located digital signal processor (DSP) resources for conferencing, transcoding, and media termination point (MTP).
- Other Unified Communications services, such as messaging (voicemail), presence, and mobility are typically co-located.

- Interfaces to legacy voice services such as PBXs and voicemail systems are connected within the campus, with no operational costs associated with bandwidth or connectivity.
- SIP-based video ISDN gateways are needed to communicate with videoconferencing devices on the public ISDN network.
- Cisco Expressway-C and Cisco Expressway-E provide a collaboration edge function that enables secure business-to-business telepresence and video communications, and enterprise access for remote and mobile workers over the internet.
- Cisco TelePresence Video Communication Server (VCS) may also be used to register legacy H.323 and third-party telepresence endpoints. However, to avoid the dial plan and call admission control complexities that dual call control introduces (see Design Considerations for Dual Call Control Deployments, page 10-40), Cisco recommends using SIP to register all TelePresence endpoints and room-based TelePresence conferencing systems with Cisco Unified CM.
- High-bandwidth audio is available (for example, G.711 or G.722) between devices within the site.
- High-bandwidth video (for example, 1.5 Mbps with 4CIF or 720p, to 2 Mbps with 1080p) is available between devices within the site.

Best Practices for the Campus Model

Follow these guidelines and best practices when implementing the single-site model:

- Ensure that the infrastructure is highly available, enabled for QoS, and configured to offer resiliency, fast convergence, and inline power.
- Know the calling patterns for your enterprise. Use the campus model if most of the calls from your enterprise are within the same site or to PSTN users outside your enterprise.
- Use G.711 codecs for all endpoints. This practice eliminates the consumption of digital signal processor (DSP) resources for transcoding, and those resources can be allocated to other functions such as conferencing and media termination points (MTPs).
- Implement the recommended network infrastructure for high availability, connectivity options for phones (in-line power), Quality of Service (QoS) mechanisms, and security. (See Network Infrastructure, page 3-1.)
- Follow the provisioning recommendations listed in the chapter on Call Processing, page 9-1.

Multisite Deployments with Centralized Call Processing

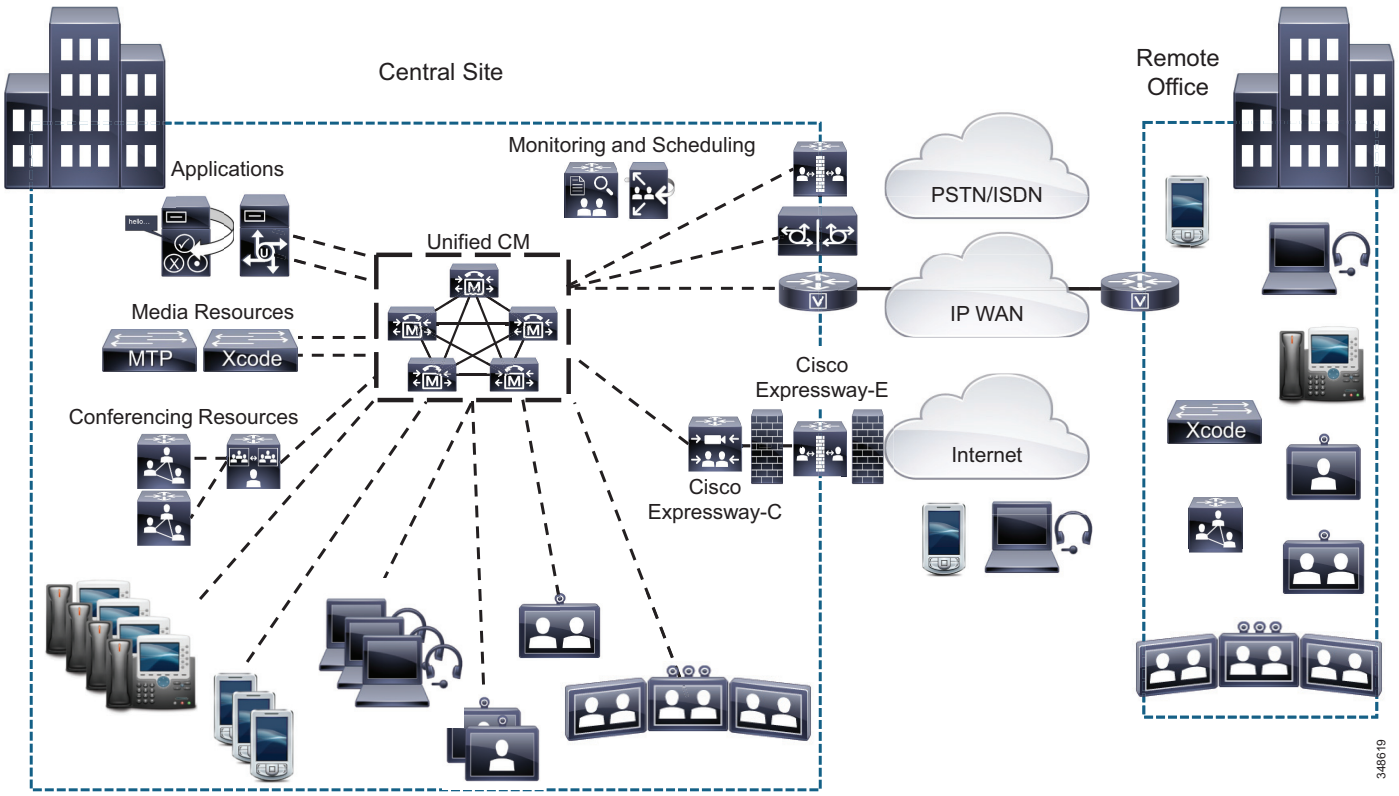
In this call processing deployment model, at least some endpoints are located remotely from the call processing service, across a QoS-enabled Wide Area Network. Due to the limited quantity of bandwidth available across the WAN, a call admission control mechanism is required to manage the number of calls admitted on any given WAN link, to keep the load within the limits of the available bandwidth. On-net communication between the endpoints traverses either a LAN/MAN (when endpoints are located in the same site) or a WAN (when endpoints are located in different sites). Communication outside the enterprise goes over an external network such as the PSTN, through a gateway or Cisco Unified Border Element (CUBE) session border controller (SBC) that can be co-located with the endpoint or at a different location (for example, when using a centralized gateway at the main site or when doing Tail End Hop Off (TEHO) across the enterprise network).

The IP WAN also carries call control signaling between the central site and the remote sites. Figure 10-3 illustrates a typical centralized call processing deployment, with a Unified CM cluster as the call processing agent at the central site and a QoS-enabled IP WAN to connect all the sites. In this deployment model, other Unified Communications services such as voice messaging, presence and

mobility are often hosted at the central site as well to reduce the overall costs of administration and maintenance. In situations where the availability of the WAN is unreliable or when WAN bandwidth costs are high, it is possible to consider decentralizing some Unified Communications services such as voice messaging (voicemail) so that the service's availability is not impacted by WAN outages.

Note In each solution for the centralized call processing model presented in this document, the various sites connect to an IP WAN with QoS enabled.

Figure 10-3 Multisite Deployment with Centralized Call Processing



The multisite model with centralized call processing has the following design characteristics:

- Single Unified CM cluster (Enterprise or Business Edition 7000). Some centralized call processing deployments may require more than one Unified CM cluster, for instance, if scale calls for more endpoints than can be serviced by a single cluster or if a cluster needs to be dedicated to an application such as a call center.
- Cisco Business Edition 6000 may be deployed in centralized call processing configurations for up to 49 remote sites.
- Cisco Business Edition 4000 may be deployed in a centralized call processing configuration.
- Maximum of 40,000 configured and registered Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) IP phones, softphones, analog ports, video endpoints, SIP-based TelePresence endpoints and room-based TelePresence conferencing systems, mobile clients, and Cisco Virtualization Experience Clients (VXC) per Unified CM cluster.

- Maximum of 2,000 locations or branch sites per Unified CM cluster.
- Maximum of 2,100 gateways and trunks (that is, the total number of H.323 gateways, H.323 trunks, digital MGCP devices, and SIP trunks) per Unified CM cluster.
- PSTN connectivity for all off-net calls.
- Digital signal processor (DSP) resources for conferencing, transcoding, and media termination point (MTP) are distributed locally to each site to reduce WAN bandwidth consumption on calls requiring DSPs.
- Multipoint control unit (MCU) or other multipoint conferencing resources are required for multipoint conferencing. These resources may all be located at the central site or may be distributed to the remote sites if local conferencing resources are required.
- Capability to integrate with legacy private branch exchange (PBX) and voicemail systems. Connections to legacy voice services such as PBXs and voicemail systems can be made within the central site, with no operational costs associated with bandwidth or connectivity. Connectivity to legacy systems located at remote sites may require the operational expenses associated with the provisioning of extra WAN bandwidth.
- SIP-based video ISDN gateways are needed to communicate with videoconferencing devices on the public ISDN network. ISDN video gateways can be centralized and/or deployed at each remote site.
- Cisco Expressway-C and Cisco Expressway-E provide a collaboration edge function that enables secure business-to-business telepresence and video communications, and VPN-less enterprise access for remote and mobile workers over the Internet.
- Cisco TelePresence Video Communication Server (VCS) may also be used to register legacy H.323 and third-party telepresence endpoints. However, to avoid the dial plan and call admission control complexities that dual call control introduces (see Design Considerations for Dual Call Control Deployments, page 10-40), Cisco recommends using SIP to register all TelePresence endpoints and room-based TelePresence conferencing systems with Cisco Unified Communications Manager
- The system allows for the automated selection of high-bandwidth audio (for example, G.711 or G.722) between devices within the site, while selecting low-bandwidth audio (for example, G.729) between devices in different sites.
- The system allows for the automated selection of high-bandwidth video (for example, 1.5 Mbps with 4CIF or 720p, to 2 Mbps with 1080p) between devices in the same site, and low-bandwidth video (for example, 384 kbps with 448p or CIF) between devices at different sites.
- A minimum of 1.5 Mbps or greater WAN link speed should be used when video is to be placed on the WAN.
- Call admission control is achieved through Enhanced Locations CAC.
- For voice and video calls, automated alternate routing (AAR) provides the automated rerouting of calls through the PSTN when call admission control denies a call between endpoints within a cluster due to lack of bandwidth. AAR relies on a gateway being available to route the call from the calling phone toward the PSTN, and another gateway to accept the call from the PSTN at the remote site, to be connected to the called phone.
- Call Forward Unregistered (CFUR) functionality provides the automated rerouting of calls through the PSTN when an endpoint is considered unregistered due to a remote WAN link failure. CFUR relies on a gateway being available to route the call from the calling phone toward the PSTN, and another gateway to accept the call from the PSTN at the remote site, to be connected to the called phone.

- Survivable Remote Site Telephony (SRST) for video endpoints located at remote sites renders all devices as audio-only if the WAN connection fails. Enhanced SRST enables video survivability on SIP video endpoints (Cisco Unified IP Phone 9900, for example) during WAN failure. For SRST video support with a particular phone model, refer to the respective Cisco Unified IP Phone Administration Guide available at https://www.cisco.com.
- Cisco Unified Communications Manager Express (Unified CME) may be used for remote site survivability (Enhanced SRST) instead of SRST.
- Cisco Unified Communications Manager Express (Unified CME) can be integrated with the Cisco Unity Connection server in the branch office or remote site. The Cisco Unity Connection server is registered to the Unified CM at the central site in normal mode and can fall back to Enhanced SRST mode when Unified CM is not reachable, or during a WAN outage, to provide the users at the branch offices with access to their voicemail with MWI.
- With multisite centralized call processing model, PSTN routing through both central and remote site gateways is supported. Providing a local gateway at a remote site for local PSTN breakout might be a requirement for countries that provide emergency services for users located at remote sites. In this case, the local gateway at the remote site provides call routing to the local PSAP for emergency calls. Local PSTN breakout at remote sites might also be required for countries having strict regulations that require the separation of the IP telephony network from the PSTN. Where regulations allow, local PSTN breakout through the remote site gateway can be used to enable toll bypass or tail-end hop off (TEHO).

Connectivity options for the IP WAN include:

- Leased lines
- Frame Relay
- Asynchronous Transfer Mode (ATM)
- ATM and Frame Relay Service Inter-Working (SIW)
- Multiprotocol Label Switching (MPLS) Virtual Private Network (VPN)
- Voice and Video Enabled IP Security Protocol (IPSec) VPN (V3PN)

Routers that reside at the WAN edges require quality of service (QoS) mechanisms, such as priority queuing and traffic shaping, to protect the voice and video traffic from the data traffic across the WAN, where bandwidth is typically scarce. In addition, a call admission control scheme is needed to avoid oversubscribing the WAN links with voice and/or video traffic and deteriorating the quality of established calls. For centralized call processing deployments, Enhanced Location CAC or RSVP-enabled locations configured within Unified CM provide call admission control (CAC). (Refer to the chapter on Bandwidth Management, page 13-1, for more information on locations.)

A variety of Cisco gateways can provide the remote sites with TDM and/or IP-based PSTN access. When the IP WAN is down, or if all the available bandwidth on the IP WAN has been consumed, calls from users at remote sites can be rerouted through the PSTN. The Cisco Unified Survivable Remote Site Telephony (SRST) feature, available for both SCCP and SIP phones, provides call processing at the branch offices for Cisco Unified IP Phones if they lose their connection to the remote primary, secondary, or tertiary Unified CM or if the WAN connection is down. Cisco Unified SRST and Cisco Unified CME with Enhanced SRST are available on Cisco IOS gateways and routers. Unified CME with Enhanced SRST provides more features for the phones than regular Unified SRST.

Best Practices for the Centralized Call Processing Model

Follow these guidelines and best practices when implementing multisite centralized call processing deployments:

- Minimize delay between Unified CM and remote locations to reduce voice cut-through delays (also known as clipping).
- Configure Enhanced Locations CAC in Unified CM to provide call admission control into and out of remote branches. See the chapter on Bandwidth Management, page 13-1, for details on how to apply this mechanism to the various WAN topologies.
- The number of IP phones and line appearances supported in Survivable Remote Site Telephony (SRST) mode at each remote site depends on the branch router platform, the amount of memory installed, and the Cisco IOS release. SRST supports up to 1,500 phones, while Unified CME running Enhanced SRST supports 450 phones. (For the latest SRST or Unified CME platform and code specifications, refer to the SRST and Unified CME documentation available at https://www.cisco.com.) Generally speaking, however, the choice of whether to adopt a centralized call processing or distributed call processing approach for a given site depends on a number of factors such as:
  - IP WAN bandwidth or delay limitations
  - Criticality of the voice network
  - Feature set needs
  - Scalability
  - Ease of management
  - Cost

  If a distributed call processing model is deemed more suitable for the customer's business needs, the choices include installing a Unified CM cluster at each site or running Unified CME at the remote sites.
- At the remote sites, use the following features to ensure call processing survivability in the event of a WAN failure:
  - For SCCP phones, use SRST or Enhanced SRST.
  - For SIP phones, use SIP SRST or Enhanced SRST.
  - For deployments with centralized voicemail, use Survivable Remote Site Voicemail (SRSV).

  SRST, Enhanced SRST, SIP SRST, SRSV, and MGCP Gateway Fallback can reside with each other on the same Cisco IOS gateway.

Remote Site Survivability

When deploying Cisco Unified Communications across a WAN with the centralized call processing model, you should take additional steps to ensure that data and voice services at the remote sites are highly available. Table 10-3 summarizes the different strategies for providing high availability at the remote sites. The choice of one of these strategies may depend on several factors, such as specific business or application requirements, the priorities associated with highly available data and voice services, and cost considerations.

Table 10-3 Strategies for High Availability at the Remote Sites

Strategy	High Availability for Data Services?	High Availability for Voice Services?
Redundant IP WAN links in branch router	Yes	Yes
Redundant branch router platforms + Redundant IP WAN links	Yes	Yes
Data-only ISDN backup + SRST or Enhanced SRST	Yes	Yes
Data and voice ISDN backup	Yes	Yes (see rules below)
Cisco Unified Survivable Remote Site Telephony (SRST) or Enhanced SRST	No	Yes

The first two solutions listed in Table 10-3 provide high availability at the network infrastructure layer by adding redundancy to the IP WAN access points, thus maintaining IP connectivity between the remote IP phones and the centralized Unified CM at all times. These solutions apply to both data and voice services, and are entirely transparent to the call processing layer. The options range from adding a redundant IP WAN link at the branch router to adding a second branch router platform with a redundant IP WAN link.

The third and forth solutions in Table 10-3 use an ISDN backup link to provide survivability during WAN failures. The two deployment options for ISDN backup are:

- Data-only ISDN backup

  With this option, ISDN is used for data survivability only, while SRST or Enhanced SRST is used for voice survivability. Note that you should configure an access control list on the branch router to prevent traffic from telephony signaling protocols such as Skinny Client Control Protocol (SCCP), H.323, Media Gateway Control Protocol (MGCP), or Session Initiation Protocol (SIP) from entering the ISDN interface, so that signaling from the IP phones does not reach the Unified CM at the central site. This is to ensure that the telephony endpoints located at the branch detect the WAN's failure and rely on local SRST resources.

- Data and voice ISDN backup

  With this option, ISDN is used for both data and voice survivability. In this case, SRST or Enhanced SRST is not used because the IP phones maintain IP connectivity to the Unified CM cluster at all times. However, Cisco recommends that you use ISDN to transport data and voice traffic only if all of the following conditions are true:

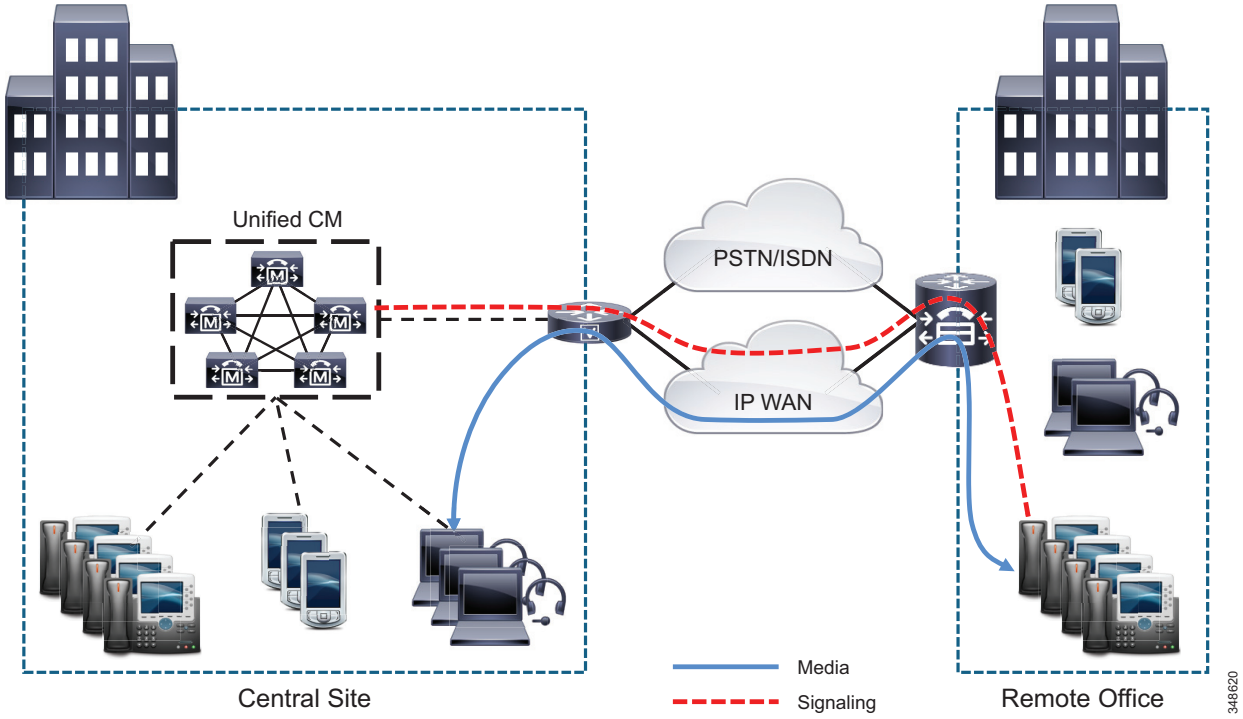
  – The bandwidth allocated to voice traffic on the ISDN link is the same as the bandwidth allocated to voice traffic on the IP WAN link.

  – The ISDN link bandwidth is fixed.

  – All the required QoS features have been deployed on the router's ISDN interfaces. Refer to the chapter on Network Infrastructure, page 3-1, for more details on QoS.

The fifth solution listed in Table 10-3, Survivable Remote Site Telephony (SRST) or Enhanced SRST, provides high availability for voice services only, by providing a subset of the call processing capabilities within the remote office router and enhancing the IP phones with the ability to “re-home” to the call processing functions in the local router if a WAN failure is detected. Figure 10-4 illustrates a typical call scenario with SRST or Enhanced SRST.

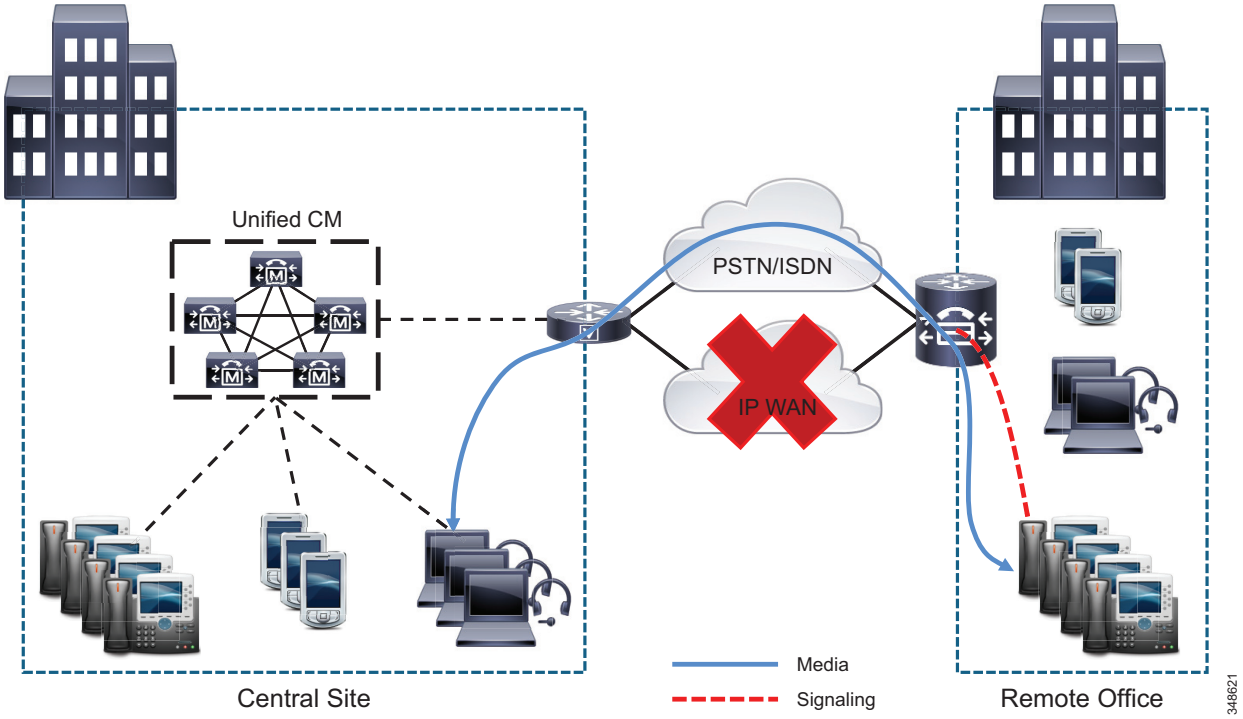
Figure 10-4 Survivable Remote Site Telephony (SRST) or Enhanced SRST, Normal Operation



Under normal operations shown in Figure 10-4, the remote office connects to the central site via an IP WAN, which carries data traffic, voice traffic, and call signaling. The IP phones at the remote office exchange call signaling information with the Unified CM cluster at the central site and place their calls across the IP WAN. The remote office router or gateway forwards both types of traffic (call signaling and voice) transparently and has no knowledge of the IP phones.

If the WAN link to the remote office fails, as shown in Figure 10-5, or if some other event causes loss of connectivity to the Unified CM cluster, the remote office IP phones re-register with the remote office router in SRST mode. The remote office router, using SRST or Enhanced SRST, queries the IP phones for their configuration and uses this information to build its own configuration automatically. The remote office IP phones can then make and receive calls either within the remote office network or through the PSTN. The phone displays the message “Unified CM fallback mode,” and some advanced Unified CM features are unavailable and are grayed out on the phone display.

Figure 10-5 Survivable Remote Site Telephony (SRST) or Enhanced SRST, WAN Failure



When WAN connectivity to the central site is reestablished, the remote office IP phones automatically re-register with the Unified CM cluster and resume normal operation. The remote office SRST router deletes its information about the IP phones and reverts to its standard routing or gateway configuration. Routers using Enhanced SRST at the remote office can choose to save the learned phone and line configuration to the running configuration on the Unified CME router by using the auto-provision option. If **auto-provision none** is configured, none of the auto-provisioned phone or line configuration information is written to the running configuration of the Unified CME router. Hence, no configuration change is required on Unified CME if the IP phone is replaced and the MAC address changes.

Note When WAN connectivity to the central site is reestablished, or when Unified CM is reachable again, phones in SRST mode with active calls will not immediately re-register to Unified CM until those active calls are terminated.

Enhanced SRST

Enhanced SRST provides more call processing features for the IP phones than are available with the SRST feature on a router. In addition to the SRST features such as call preservation, auto-provisioning, and failover, Enhanced SRST also provides most of the Unified CME telephony features for phones, including:

- Paging
- Conferencing
- Hunt groups
- Basic automatic call distribution (B-ACD)

- Call park, call pickup, call pickup groups
- Overlay-DN, softkey templates
- Cisco IP Communicator
- Cisco Jabber Clients
- Cisco Unified Video Advantage
- Endpoint video calls

Enhanced SRST provides call processing support for SCCP and SIP phones in case of a WAN failure. However, Enhanced SRST does not provide fallback support for MGCP phones or endpoints. To enable MGCP phones to fall back if they lose their connection to Unified CM, or if the WAN connection fails, you can additionally configure the MGCP Gateway Fallback feature on the same Unified CME server running as the SRST fallback server.

Best Practices for Enhanced SRST

- Use the Unified CME IP address as the IP address for SRST reference in the Unified CM configuration.
- The Connection Monitor Duration is a timer that specifies how long phones monitor the WAN link before initiating a fallback from SRST to Unified CM. The default setting of 120 seconds should be used in most cases. However, to prevent phones in SRST mode from falling back and re-homing to Unified CM with flapping links, you can set the Connection Monitor Duration parameter on Unified CM to a longer period so that phones do not keep registering back and forth between the SRST router and Unified CM. Do not set the value to an extensively longer period because this will prevent the phones from falling back from SRST to Unified CM for a long amount of time.
- Phones in SRST fallback mode will not re-home to Unified CM when they are in active state.
- Phones in SRST fallback mode revert to non-secure mode from secure conferencing.
- Configure **auto-provision none** to prevent writing any learned ephone-dn or ephone configuration to the running configuration of the Unified CME router. This eliminates the need to change the configuration if the IP phone is replaced or the MAC address changes.

For more information on Enhanced SRST, refer to the *Cisco Unified Communications Manager Express System Administrator Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-express/products-installation-and-configuration-guides-list.html

For more information on MGCP Gateway fallback, refer to the information on MGCP gateway fallback in the *Cisco Unified Communications Manager and Interoperability Configuration Guide, Cisco IOS Release 15M&T*, available at

https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/voice/cminterop/configuration/15-mt/cminterop-15-mt-book.html

Best Practices for SRST Router

Use a Cisco Unified SRST router, rather than Enhanced SRST, for the following deployment scenarios:

- For supporting a maximum of 1,500 phones on a single SRST router. (Enhanced SRST supports a maximum of 450 phones.)
- For up to 3,000 phones, use two SRST routers. Dial plans must be properly configured to route the calls back and forth between the SRST routers.

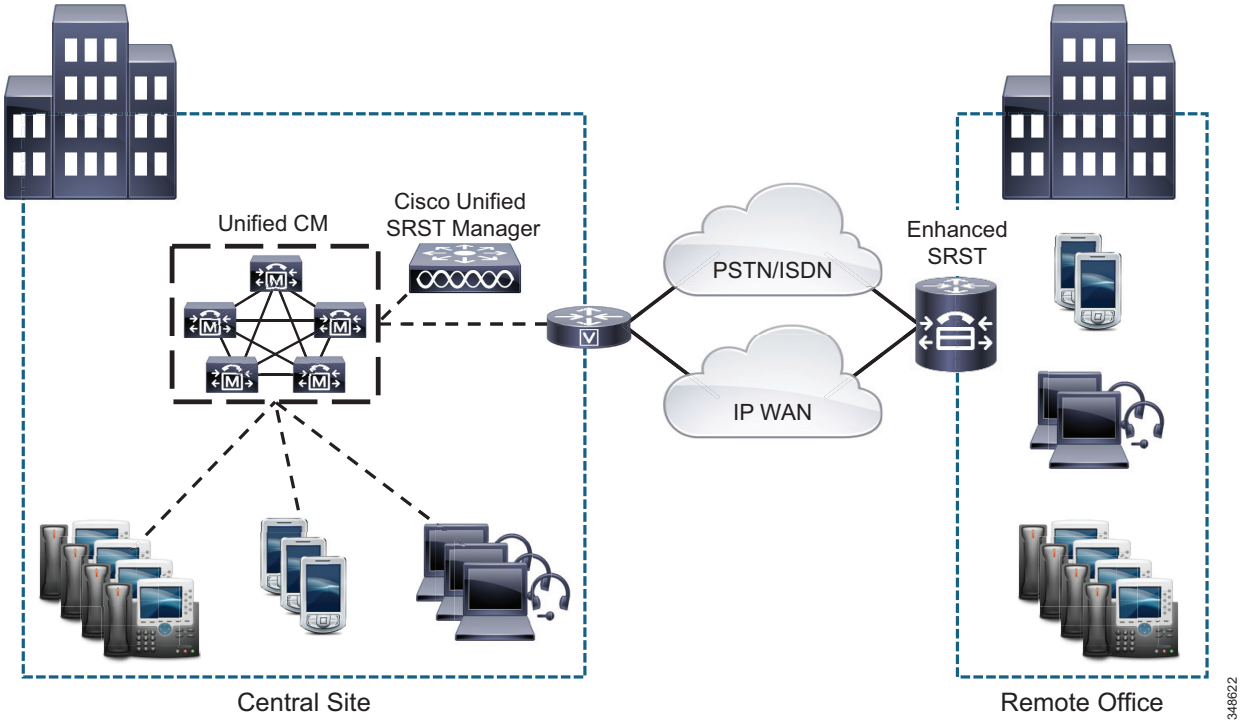
- For simple, one-time configuration of basic SRST functions.
- For SRTP media encryption, which is available only in Cisco Unified SRST (Secure SRST).

For routing calls to and from phones that are unreachable or not registered to the SRST router, use the **alias** command.

Cisco Unified Survivable Remote Site Telephony Manager

Cisco Unified Survivable Remote Site Telephony (SRST) Manager simplifies the deployment of Enhanced SRST as well as traditional SRST in the branch. (See Figure 10-6.) Cisco Unified SRST Manager is Linux-based software running inside a virtual machine on Cisco supported virtualized platforms (for example, Cisco UCS). Cisco Unified SRST Manager supports only the centralized call processing deployment model, where the Cisco Unified CM cluster runs in the central location. Cisco Unified SRST Manager can be deployed in the central location along with the Cisco Unified CM cluster or in the remote branch location. Figure 10-6 illustrates the deployment of Cisco Unified SRST Manager in the central location. During normal operation, Cisco Unified SRST Manager regularly retrieves configurations (for example, calling search space, partition, hunt group, call park, call pickup, and so forth, if configured) from Cisco Unified CM and uploads them to provision the branch router with similar functionality for use in SRST mode. Thus, Cisco Unified SRST Manager reduces manual configuration required in the branch SRST router and enables users to have a similar calling experience in both SRST and normal modes.

Figure 10-6 Cisco Unified SRST Manager Deployed in the Central Location



Cisco Unified SRST Manager consumes bandwidth from the WAN link when uploading the Unified CM configurations to provision the remote office router. The Cisco Unified SRST Manager software does not perform packet marking, therefore the Cisco Unified SRST Manager traffic will travel as best-effort on the network. Cisco recommends maintaining this best-effort marking, which is IP Precedence 0

(DSCP 0 or PHB BE), to ensure that it does not interfere with real-time high priority voice traffic. To ensure that Cisco Unified SRST Manager traffic does not cause congestion and to reduce the chances of packet drop, Cisco recommends scheduling the configuration upload to take place during non-peak hours (for example, in the evening hours or during the weekend). The configuration upload schedule can be set from the Cisco Unified SRST Manager web interface.

Consider the following guidelines when you deploy Cisco Unified SRST Manager:

- Cisco Unified SRST Manager is not supported with the Cisco Unified Communications 500 Series platform.
- The remote office voice gateway must be co-resident with (reside on) the SRST router.
- There is no high availability support with Cisco Unified SRST Manager. If Cisco Unified SRST Manager is unavailable, configuration upload is not possible.
- Cisco Unified SRST Manager is not supported in deployments where NAT is used between the headquarters and branch locations.

Voice over the PSTN as a Variant of Centralized Call Processing

Centralized call processing deployments can be adapted so that inter-site voice media is sent over the PSTN instead of the WAN. With this configuration, the signaling (call control) of all telephony endpoints is still controlled by the central Unified CM cluster, therefore this Voice over the PSTN (VoPSTN) model variation still requires a QoS-enabled WAN with appropriate bandwidth configured for the signaling traffic.

VoPSTN can be an attractive option in deployments where IP WAN bandwidth is either scarce or expensive with respect to PSTN charges, or where IP WAN bandwidth upgrades are planned for a later date but the Cisco Unified Communications system is already being deployed.

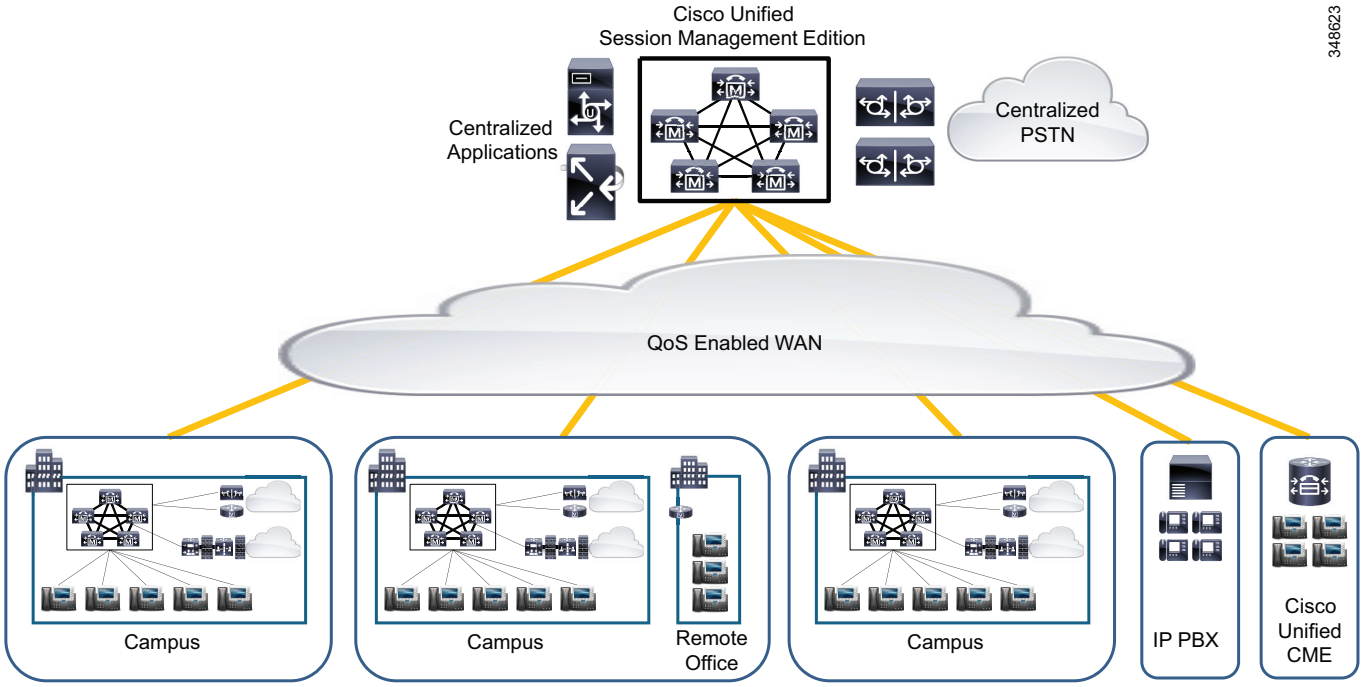
For more information on VoPSTN deployment options and design guidance, refer to the VoPSTN sections in the *Unified Communications Deployment Models* chapter of the *Cisco Unified Communications System 9.0 SRND*, available at

https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cucm/srnd/9x/uc9x/models.html

Multisite Deployments with Distributed Call Processing

The model for a multisite deployment with distributed call processing consists of multiple independent sites, each with its own call processing agent cluster connected to an IP WAN that carries voice traffic between the distributed sites. Figure 10-7 illustrates a typical distributed call processing deployment.

Figure 10-7 Multisite Deployment with Distributed Call Processing



Each site in the distributed call processing model can be one of the following:

- A single site with its own call processing agent, which can be either:
  - Cisco Unified Communications Manager (Enterprise or Business Edition 7000)
  - Cisco Business Edition 6000
  - Cisco Unified Communications Manager Express (Unified CME)
  - A third-party IP PBX
  - A legacy PBX with Voice over IP (VoIP) gateway
- A centralized call processing site and all of its associated remote sites

The multisite model with distributed call processing has the following design characteristics:

- A centralized platform for trunk and dial plan aggregation is commonly deployed. This platform is typically a Cisco Unified Communications Manager Session Management Edition (SME) cluster, although a Session Initiation Protocol (SIP) Proxy Server could also be used to provide intercluster call routing and dial plan aggregation in multisite distributed call processing deployments.
- Centralized services such as:
  - Centralized PSTN access
  - Centralized voicemail
  - Centralized conferencing

  These services can be deployed centrally, thus benefiting from centralized management and economies of scale. Services that need to track end-user status (for example, Cisco IM and Presence) must connect to the Unified CM cluster for the users that they serve.
- High-bandwidth audio (for example, G.711 or G.722) between devices in the same site, but low-bandwidth audio (for example, G.729) between devices in different sites.
- High-bandwidth video (for example, 1.5 Mbps with 4CIF or 720p, to 2 Mbps with 1080p) between devices in the same site, but low-bandwidth video (for example, 384 kbps with 448p or CIF) between devices at different sites.
- Minimum of 1.5 Mbps or greater WAN link speeds. Video is *not* recommended on WAN connections that operate at speeds lower than 1.5 Mbps.
- Call admission control is achieved through Enhanced Locations CAC.

An IP WAN interconnects all the distributed call processing sites. Typically, the PSTN serves as a backup connection between the sites in case the IP WAN connection fails or does not have any more available bandwidth. A site connected only through the PSTN is a standalone site and is not covered by the distributed call processing model. (See Campus Deployments, page 10-10.)

Connectivity options for the IP WAN include:

- Leased lines
- Frame Relay
- Asynchronous Transfer Mode (ATM)
- ATM and Frame Relay Service Inter-Working (SIW)
- Multiprotocol Label Switching (MPLS) Virtual Private Network (VPN)
- Voice and Video Enabled IP Security Protocol (IPSec) VPN (V3PN)

Best Practices for the Distributed Call Processing Model

A multisite deployment with distributed call processing has many of the same requirements as a single site or a multisite deployment with centralized call processing. Follow the best practices from these other models in addition to the ones listed here for the distributed call processing model. (See Campus Deployments, page 10-10, and Multisite Deployments with Centralized Call Processing, page 10-12.)

Dial Plan Aggregation Platforms for Distributed Call Processing Deployments

A Cisco Unified Communications Manager Session Management Edition (SME) cluster or Session Initiation Protocol (SIP) proxy servers can be used to provide intercluster call routing and dial plan aggregation in multisite distributed call processing deployments. The following best practices apply to the use of these trunk and dial plan aggregation devices:

Unified CM Session Management Edition Clusters

Cisco Unified Communications Manager Session Management Edition is commonly used for intercluster call routing and dial plan aggregation in distributed call processing deployments. Intercluster call routing can be number based using standard numeric route patterns, or URI and number based using the Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR) (see Global Dial Plan Replication, page 14-47). Unified CM Session Management Edition uses exactly the same code and user interface as Unified CM but leverages support for multiple trunk protocols (SIP, H.323, and MGCP) as well as sophisticated trunk, digit manipulation, and call admission control features. Unified CM Session Management Edition cluster deployments typically consist of many trunks (SIP trunks are recommended; see Cisco Unified CM Trunks, page 6-1) and no Unified Communications endpoints. Unified CM Session Management Edition clusters can use all of the high availability features (such as clustering over the WAN, and Run on all Unified CM Nodes) that are available to Unified CM clusters.

SIP Proxy Deployments

SIP proxies such as the Cisco Unified SIP Proxy provide call routing and SIP signaling normalization.

The following best practices apply to the use of SIP proxies:

- Provide adequate redundancy for the SIP proxies.
- Ensure that the SIP proxies have the capacity for the call rate and number of calls required in the network.

Note Because Session Management Edition (SME) uses exactly the same code and GUI as Unified CM and can also share intercluster features such as ILS, GDPR, and Enhanced Locations Call Admission Control (ELCAC), SME is the recommended trunk and dial plan aggregation platform in multi-site distributed call processing deployments.

Leaf Unified Communications Systems for the Distributed Call Processing Model

Your choice of call processing agent will vary, based on many factors. The main factors, for the purpose of design, are the size of the site and the functionality required.

For a distributed call processing deployment, each site may have its own call processing agent. The design of each site varies with the call processing agent, the functionality required, and the fault tolerance required. For example, in a site with 500 phones, a Unified CM cluster containing two servers can provide one-to-one redundancy, with the backup server being used as a publisher and Trivial File Transfer Protocol (TFTP) server.

The requirement for IP-based applications also greatly affects the choice of call processing agent because only Unified CM provides the required support for many Cisco IP applications.

Table 10-4 lists recommended call processing agents.

Table 10-4 Recommended Call Processing Agents

Call Processing Agent	Recommended Size	Comments
Cisco Unified Communications Manager Express (Unified CME)	Up to 450 phones	• For small remote sites • Capacity depends on Cisco IOS platform • SIP trunks are recommended
Cisco Business Edition 6000	Up to 2,500 phones	• For small to medium sites • Supports centralized call processing • Supports distributed call processing • SIP trunks are recommended
Cisco Business Edition 4000	Up to 200 phones	• For small sites • Supports centralized call processing
Cisco Unified Communications Manager (Enterprise or Business Edition 7000)	50 to 40,000 phones	• Small to large sites, depending on the size of the Unified CM cluster • Supports centralized call processing • Supports distributed call processing • SIP trunks are recommended
IP PBX	Depends on the PBX	• IP PBXs commonly use SIP trunks, which can be used to connect to SME
Legacy PBX with VoIP gateway	Depends on the PBX	• Number of IP WAN calls and functionality depend on the PBX-to-VoIP gateway protocol and the gateway platform • SIP trunks are recommended between the VoIP gateway and SME

Unified CM Session Management Edition

Cisco Unified CM Session Management Edition (SME) is the recommended trunk and dial plan aggregation platform in multi-site distributed call processing deployments. SME is essentially a Unified CM cluster with trunk interfaces only and no IP endpoints. It enables aggregation of multiple unified communications systems, referred to as leaf systems.

Cisco Unified CM Session Management Edition supports the following trunk protocols:

- SIP intercluster trunks
- SIP trunks
- H.323 Annex M1 intercluster trunks
- H.323 trunks to gateways
- MGCP trunks to gateways

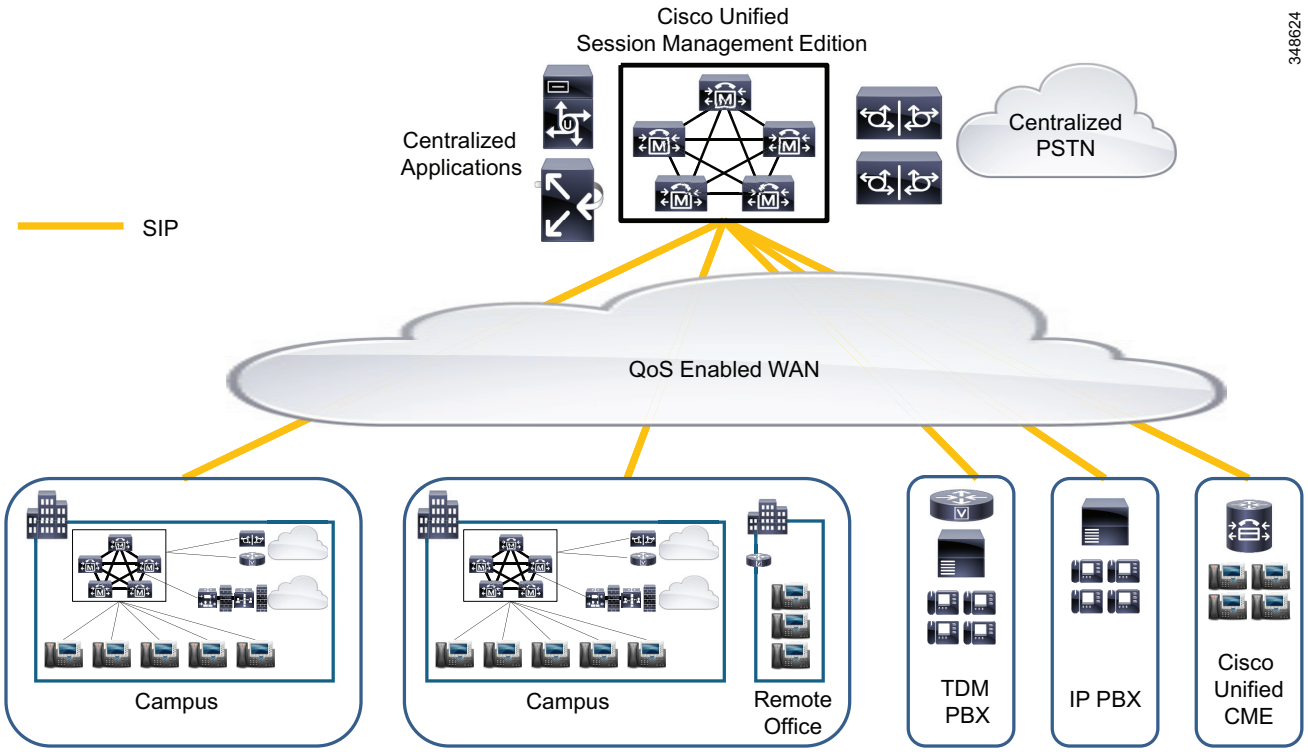
SIP trunks are recommended for SME and leaf Unified Communications systems because SIP offers additional features and functionality over H.323 and MGCP trunks. (For more information, see the chapter on Cisco Unified CM Trunks, page 6-1.)

Cisco Unified CM Session Management Edition supports the following call types:

- Voice calls
- Video calls
- Encrypted calls
- Fax calls

Unified CM Session Management Edition may also be used to connect to the PSTN and third-party unified communications systems such as PBXs and centralized unified communications applications. (See Figure 10-8.) As with any standard Unified CM cluster, third-party connections to Unified CM Session Management Edition should be system tested for interoperability prior to use in a production environment.

Figure 10-8 Multisite Distributed Call Processing Deployment with Unified CM Session Management Edition



When to Deploy Unified CM Session Management Edition

Cisco recommends deploying Unified CM Session Management Edition (SME) if you want to do any of the following:

- Create and manage a centralized dial plan

  Rather than configuring each unified communications system with a separate dial plan and trunks to connect to all the other unified communications systems, Unified CM Session Management Edition allows you to configure the leaf unified communications systems with a simplified dial plan and trunk(s) pointing to the Session Management cluster. Unified CM Session Management Edition holds the centralized dial plan and corresponding reachability information about all the other unified communications systems.

  Note Running Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR) on SME and Unified CM leaf clusters further simplifies dial plan administration because individual directory numbers, E.164 numbers corresponding to DNs, route patterns (for internal and external number ranges), and URIs can be distributed using the ILS service. This approach simplifies dial plan administration by reducing the required number of route patterns to one SIP route pattern per call control system (Unified CM cluster, for example), instead of a route pattern for each unique number range. For more information on ILS and GDPR, see Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR), page 10-32.

- Provide centralized PSTN access

  Unified CM Session Management Edition can be used to aggregate PSTN access to one (or more) centralized PSTN trunks. Centralized PSTN access is commonly combined with the reduction, or elimination, of branch-based PSTN circuits.

- Centralize applications

  The deployment of a Unified CM Session Management Edition enables commonly used applications such as conferencing or voicemail to connect directly to the Session Management cluster, thus reducing the overhead of managing multiple trunks to leaf systems.

- Aggregate PBXs for migration to a Unified Communications system

  Unified CM Session Management Edition can provide an aggregation point for multiple PBXs as part of the migration from legacy PBXs to a Cisco Unified Communications System. If ILS GDPR is deployed, the number ranges and/or URIs supported by each third-party system can also be imported into ILS GDPR and reached through a SIP route pattern and corresponding SIP trunk.

Differences Between Unified CM Session Management Edition and Standard Unified CM Clusters

The Unified CM Session Management Edition software is exactly the same as Unified CM. Unified CM Session Management Edition is designed to support a large number of trunk-to-trunk connections, and as such it is subject to the following design considerations:

Capacity

It is important to correctly size the Unified CM Session Management cluster based on the expected BHCA traffic load between leaf Unified Communications systems (for example, between Unified CM clusters and PBXs), to and from any centralized PSTN connections, and to any centralized applications. Determine the average BHCA and Call Holding Time for users of your Unified Communications system

and share this information with your Cisco account Systems Engineer (SE) or Cisco Partner to size your Unified CM Session Management Edition cluster correctly. For more information on SME sizing, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Trunks

Although SME supports SIP, H.323, and MGCP trunks, Cisco recommends using SIP as the trunk protocol for SME and Unified CM leaf clusters running Cisco Unified CM 8.5 and later releases.

SIP trunks provide a number of unique features that greatly simplify trunk designs and Unified Communications deployments, such as:

- Run on All Unified CM Nodes
- OPTIONS Ping
- Accept Codec Preference in Received Offer
- Lua scripts, which allow SIP Message and Session Description Protocol (SDP) content modification for interoperability

Using only SIP trunks in the SME cluster allows you to deploy a "media transparent" cluster where media resources, when required, are inserted by the end or leaf Unified Communications system and never by SME. Using only SIP trunks also allows you to use extended round trip times (RTTs) between SME nodes when clustering over the WAN.

Both leaf Unified CM cluster SIP trunks and SME SIP trunks should be configured as **Best Effort Early Offer** trunks. For more details on SIP trunks and **Best Effort Early Offer**, see the chapter on Cisco Unified CM Trunks, page 6-1.

Media Resources

When a media resource such as an MTP or transcoder is needed to allow a call to proceed successfully, these resources should ideally be allocated by the leaf Unified Communications systems. If SME trunk media resources are used for a call traversing the SME cluster, the media path call will hairpin through the SME media resource. By using SIP trunks only and either **Best Effort Early Offer** or **MTP-less Early Offer**, you can deploy an SME cluster without media resources. If or when media resources are required, they can be allocated by the leaf Unified Communications system.

Clustering over the WAN

SME deployments can support extended round-trip times (RTTs) of up to 500 ms between SME cluster nodes. (See Figure 10-9.) This extended RTT applies only to SME clusters (80 ms is the maximum RTT for a standard Unified CM cluster designs) and is subject to the following design restrictions:

- Extended round-trip times for SME deployments with clustering over the WAN are supported where only SIP trunks are configured in the SME cluster. All SIP trunks must be configured as either **Best Effort Early Offer** or **MTP-less Early Offer** and must use the **Run on all Unified CM Nodes** feature so that calls are not routed between nodes within the SME cluster. (For more information, see the chapter on Cisco Unified CM Trunks, page 6-1.) MGCP, SCCP, and H.323 protocols do not support extended round-trip times for SME deployments with clustering over the WAN.
- No endpoints or CTI devices are configured or registered to the SME cluster.
- No media resources such as MTPs, Trusted Relay Points (TRPs), RSVP agents, or transcoders are configured or registered to the SME cluster. (To disable media resources hosted on Unified CM nodes, deactivate the IPVMS service on each node within the cluster.)
- A minimum of 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) traffic between sites.

- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, a minimum of 1.544 Mbps (T1) bandwidth is required for database and other inter-server traffic between the publisher and every remote subscriber node.

  Like all other SME designs, your SME design must be reviewed and approved by the Cisco SME team prior to deployment.

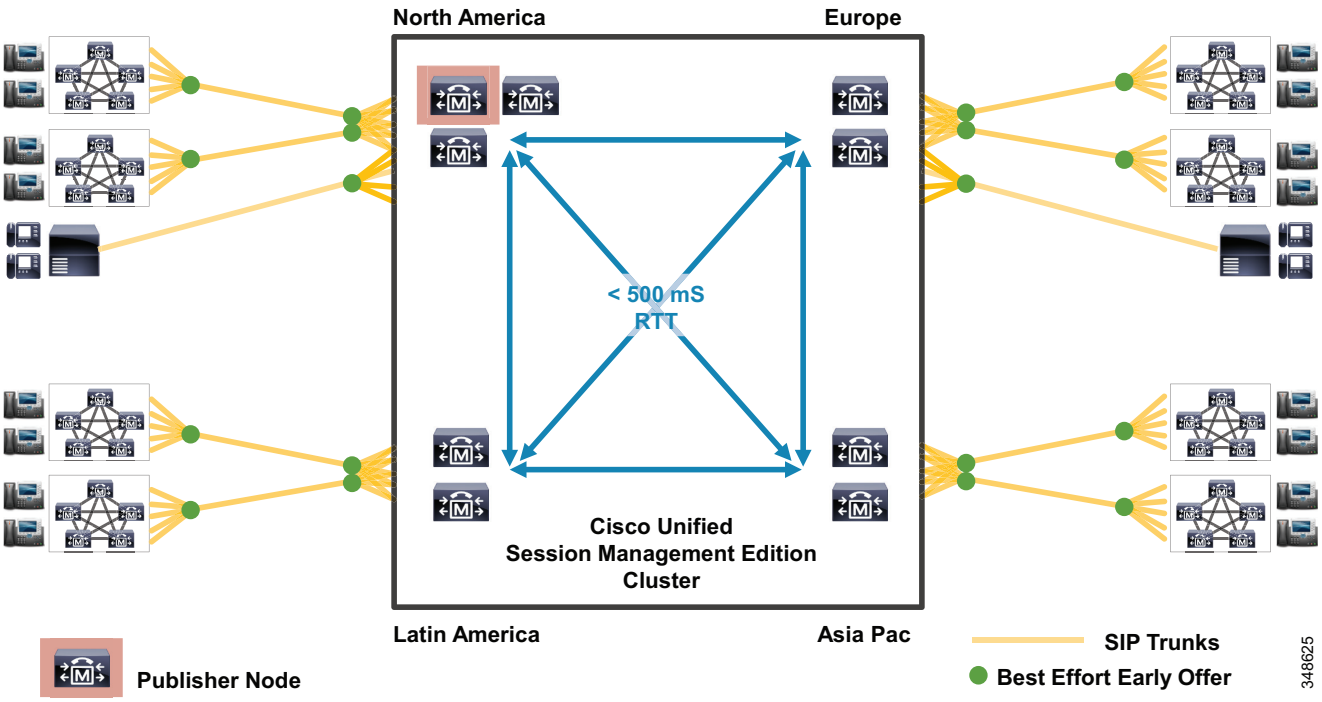
Note The upgrade process for an SME cluster consists of two key parts: Version switch-over, where the call processing node is re-booted and initialized with the new software version (this takes approximately 45 minutes per server), and database replication, where the subscriber's database is synchronized with that of the publisher node. The time taken to complete this database replication phase depends on the RTT between the publisher and subscriber nodes and the number of subscribers in the cluster. The database replication process has a minimal impact of the subscriber's call processing capability and typically can be run as a background process during normal SME cluster operation. Avoid making changes to the SME cluster configuration during the database replication phase because this increases the time it takes to complete the replication.

For SME clusters deployed with extended RTTs, before upgrading the cluster, run the following Admin level CLI command on the publisher node:

```
utils dbreplication setprocess 40
```

This command improves replication setup performance and reduces database replication times.

Figure 10-9 Unified CM Session Management Edition Clustering over the WAN with Extended Round Trip Times



Unified CM Versions

Using the latest Cisco Unified Communications System release and SIP trunks across all Unified CM leaf clusters and the SME cluster allows your Unified Communications deployment to benefit from common cross-cluster features such as Codec Preference Lists, Intercluster Lookup Service (ILS), Global Dial Plan Replication (GDPR), and Enhanced Locations Call Admission Control (ELCAC). If you do not wish to upgrade to the latest Unified Communications version on all clusters, the lowest recommended version is Cisco Unified CM 8.5 using SIP trunks, because this version includes features that improve and simplify call routing through Unified CM and Session Management Edition clusters.

Interoperability

Even though most vendors do conform to standards, differences can and do exist between protocol implementations from various vendors. As with any standard Unified CM cluster, Cisco strongly recommends that you conduct end-to-end system interoperability testing with any unverified third-party unified communications system before deploying the system in a production environment. The interoperability testing should verify call flows and features from Cisco and third-party leaf systems through the Unified CM Session Management cluster. To learn which third-party unified communications systems have been tested by the Cisco Interoperability team, refer to the information available on the Cisco Interoperability Portal at

https://www.cisco.com/c/en/us/solutions/enterprise/interoperability-portal/interOp_ucSessionMgr.html

For SIP trunk interoperability issues, Lua scripting can be used to modify inbound and outbound SIP messages and SDP content.

Load Balancing for Inbound and Outbound Calls

Configure trunks on the Unified CM Session Management Edition and leaf unified communications systems so that inbound and outbound calls are evenly distributed across the Unified CM servers within the Session Management cluster. As a general rule, always enable the **Run on All Unified CM Nodes** feature if it is available. For more information on load balancing for trunk calls, refer to the chapter on Cisco Unified CM Trunks, page 6-1.

Design Guidance and Assistance

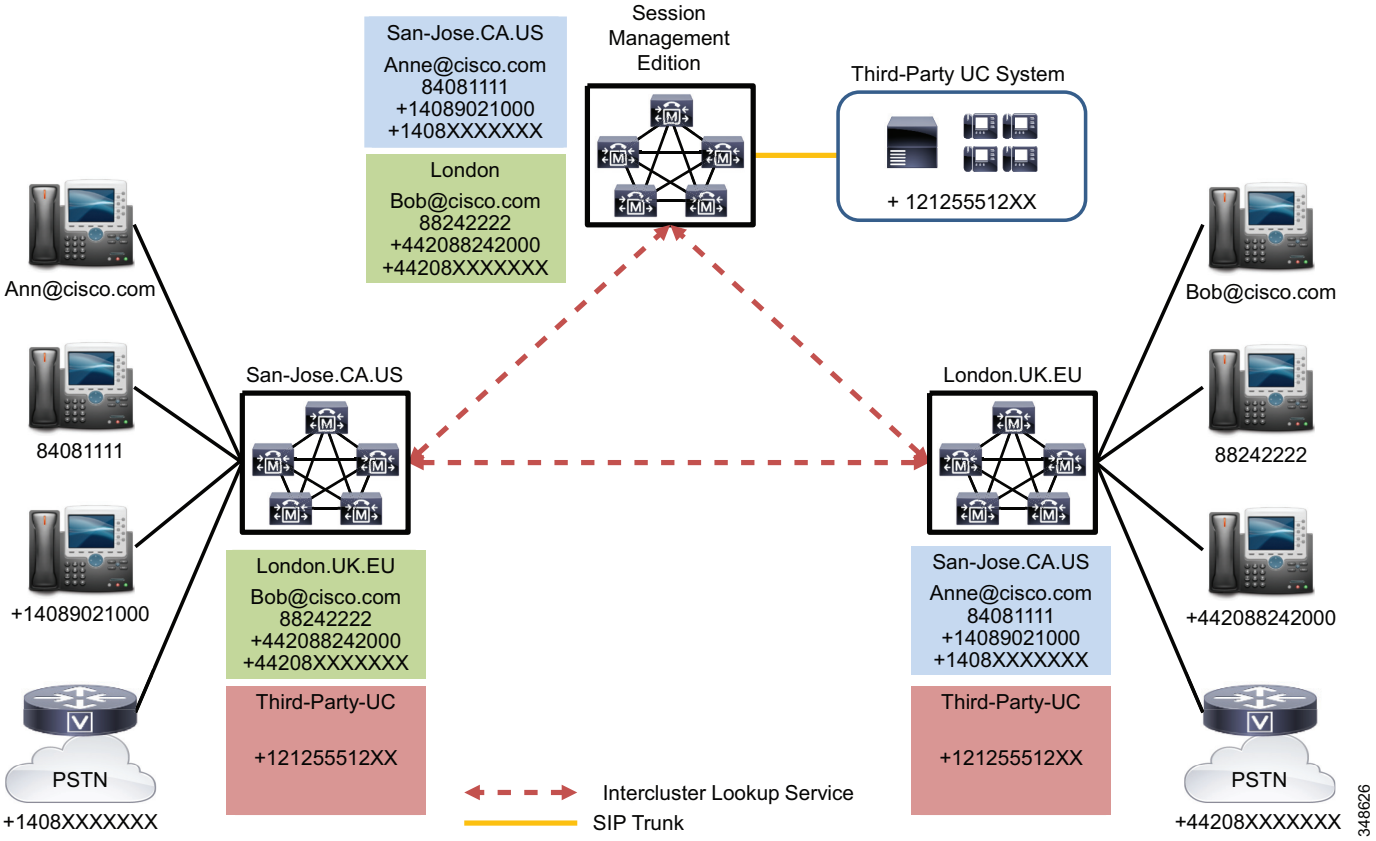
For detailed information on trunk configuration for Unified CM Session Management Edition designs and deployments, refer to the chapter on Cisco Unified CM Trunks, page 6-1.

Note Before deployment, Unified CM Session Management Edition designs should be reviewed by your Cisco SE in conjunction with the Cisco Unified CM Session Management Team.

Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR)

Global Dial Plan Replication (GDPR) uses the Intercluster Lookup Service (ILS) to share dial plan information between participating ILS-enabled clusters. GDPR allows each cluster to distribute information about its associated URIs, +E.164 numbers, enterprise numbers, +E.164 patterns, enterprise patterns, and PSTN failover numbers. Each participating cluster shares a common Global Dial Plan catalogue, which contains every number and URI advertised with GDPR and a corresponding route string that identifies in which cluster (or end Unified Communications system) the number or URI resides. (See Figure 10-10.)

Figure 10-10 ILS and GDPR Number, Pattern, and URI Distribution



With GDPR, each cluster advertises its dial plan information (numbers and URIs) with a location attribute known as a *route string*. When a call is placed to an alphanumeric URI, Unified CM checks to see whether the URI is associated to a device within the cluster. If it is not, Unified CM searches its GDPR catalogue for the URI. If a match is found in the Global Dial Plan catalogue, GDPR returns the route string that corresponds to the cluster where the number or URI resides. Unified CM uses the returned route string as a candidate to match to an existing SIP route pattern and corresponding SIP trunk. For a numeric destination, if best-match digit analysis returns a match to a destination learned via GDPR, then again the route string corresponding to the cluster where the learned destination resides is used to determine which SIP trunk to route the call to. (See Figure 10-11 and Figure 10-12.)

Figure 10-11 ILS and GDPR Number, Pattern, and URI Lookup

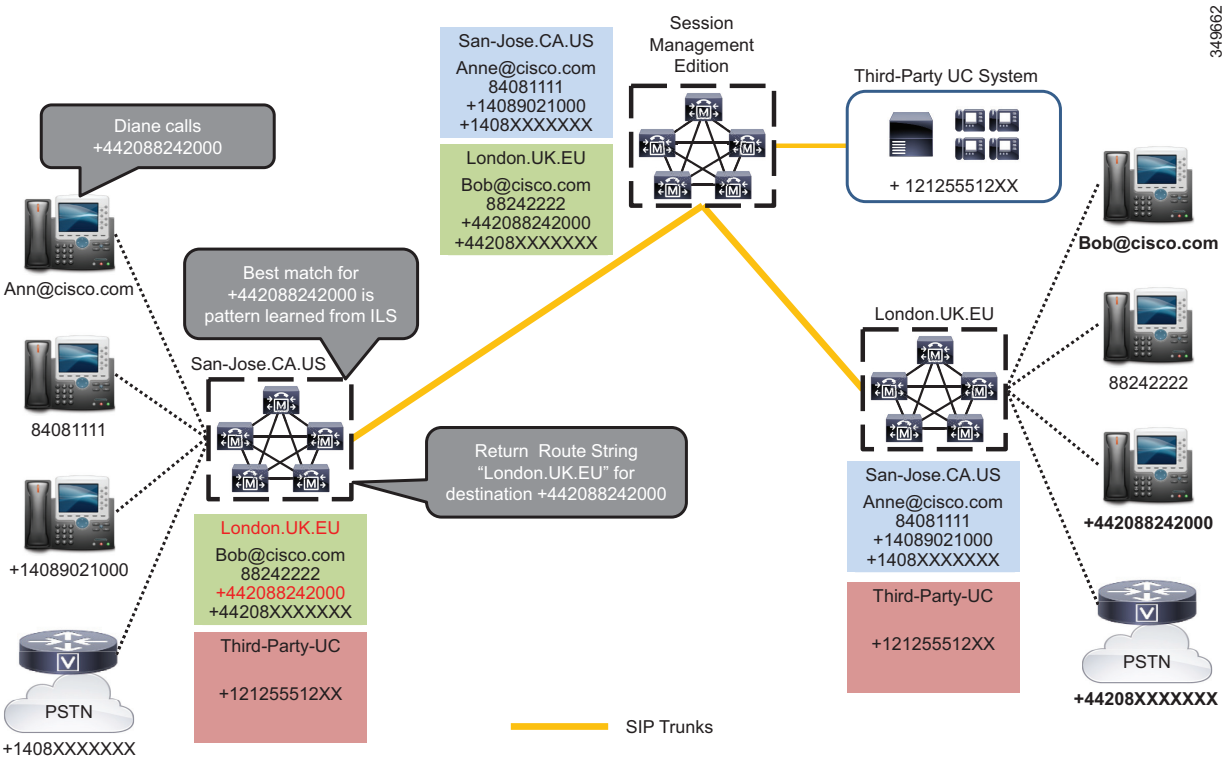
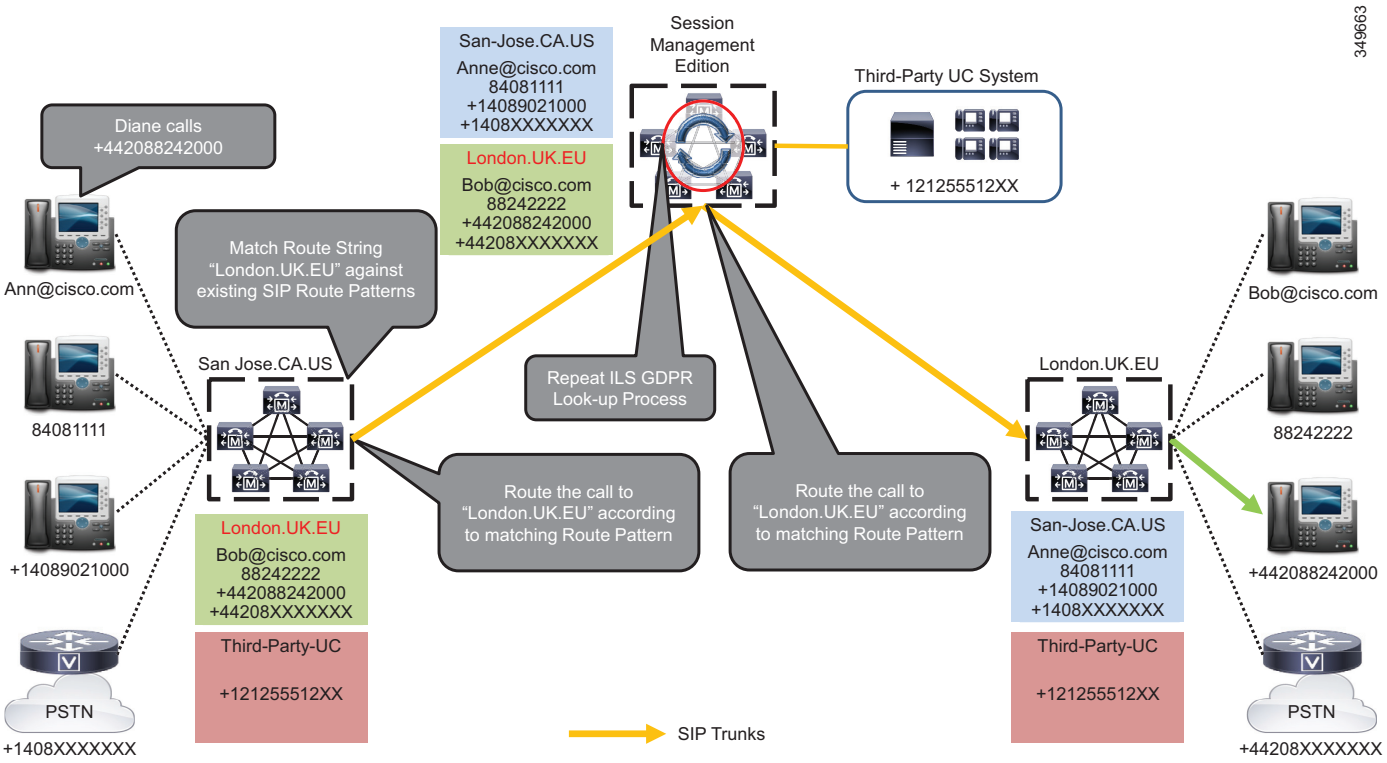


Figure 10-12 ILS and GDPR Call Routing



Benefits of ILS and GDPR

Using GDPR is significantly different from using a standard dial plan with numeric route patterns. Instead of requiring a route pattern for each unique number range within the Unified Communications network, GDPR distributes the numbers, number patterns, and URIs, and only a single SIP route pattern is required for each cluster within the Unified Communications network. Numbers and URIs associated with third-party unified communications systems (and Unified CM clusters that do not support ILS and GDPR) can be imported as catalogues into GDPR and distributed through ILS with a route string that corresponds to each unified communications system. Because both individual numbers and route patterns corresponding to groups of numbers can be advertised with GDPR, this abstraction of numbers and number ranges away from numeric route patterns allows GDPR to simply and easily support highly fragmented dial plans with many number ranges. Each cluster using ILS and GDPR can block and purge individual numbers and number ranges advertised from other participating clusters.

Each GDPR number type (+E.164 number, enterprise number, +E.164 pattern, or enterprise pattern) is placed into a specific partition when learned through ILS, allowing per-user or per-device class of service to be applied based on number type partitions and calling search spaces.

Cisco Unified Border Element also supports number and URI call routing using dial peers that match on a GDPR route string value, which is sent to Cisco Unified Border Element during call setup over a Unified CM SIP trunk. GDPR route string matching with Cisco IOS dial peers is supported with Cisco IOS releases 15.3(3)M, 15.4(1)T (ISR), 15.3(3)S (ASR), and later releases.

Deployments for the Collaboration Edge

The border between an enterprise Unified Communications network and the outside world is often referred to as the *collaboration edge*. Access to an enterprise network from the outside world can take a number of forms. For example, users can be teleworkers working from home, mobile workers with Wi-Fi internet access to the enterprise, or users making calls to and from the IP PSTN or calls to and from other businesses over the internet. The Unified Communications equipment needed at the collaboration edge largely depends upon the type of enterprise access required, which can be classified broadly into three categories:

- VPN based access
- VPN-less access
- Business-to-business communications
- IP PSTN access

These four deployment options for the collaboration edge are discussed in the sections that follow.

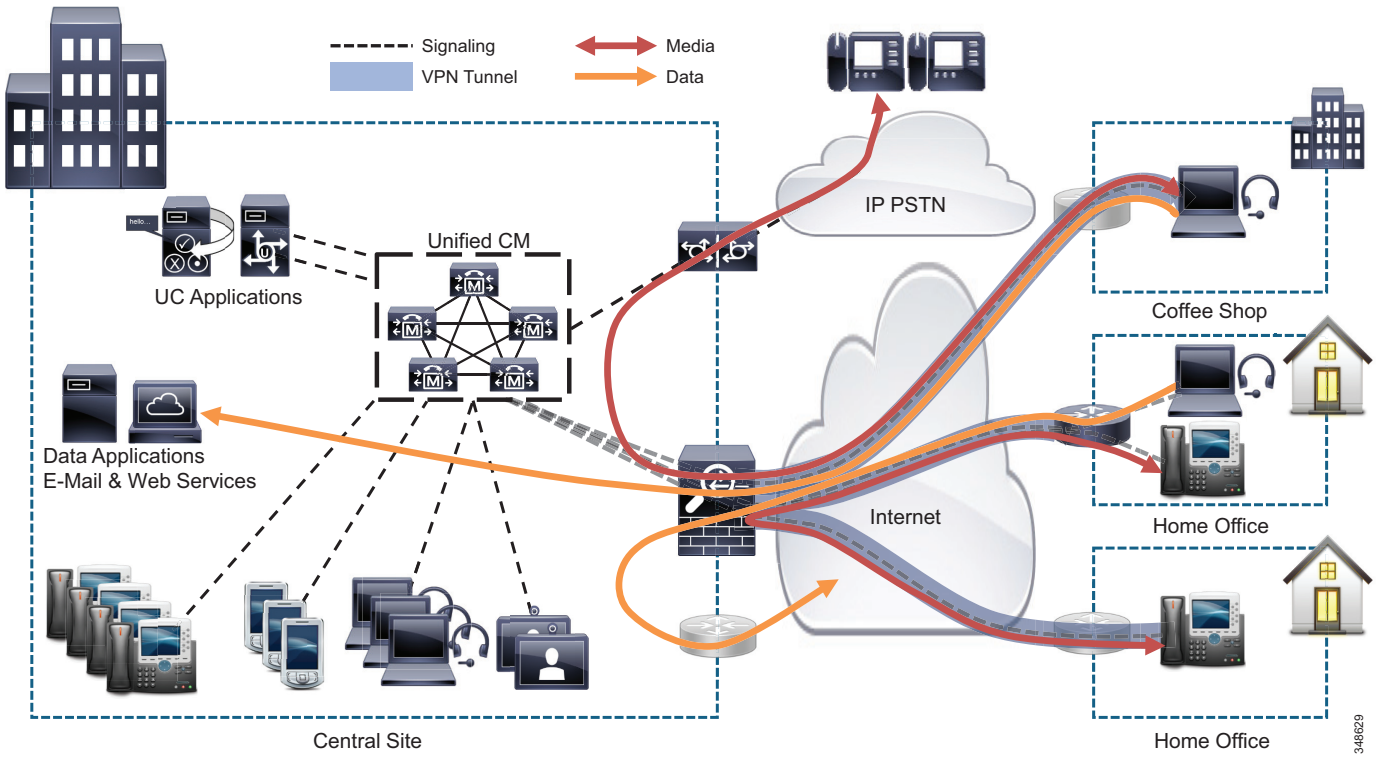
VPN Based Enterprise Access Deployments

VPN access to the enterprise network is probably the most common form of enterprise access today, and it can be provided in several ways:

- Mobile devices such as laptops, tablets, and smartphones can deploy the Cisco AnyConnect VPN client to access Unified Communications services (for example, Unified CM, Cisco IM and Presence, Cisco Unity, and others) as well as business application services (for example, the corporate email system and internal websites) within the enterprise network. With this VPN connection established, Unified Communications soft clients such as Cisco Jabber and Cisco IP Communicator can register with Unified CM and make voice, video, and encrypted calls between enterprise devices.
- Home office workers with one or more enterprise devices can deploy a Cisco Virtual Office (CVO) Integrated Services Router (ISR) to extend the enterprise network to their homes over a VPN. The CVO VPN connection provides connected devices with access to Unified Communications services (for example, Unified CM, Cisco IM and Presence, Cisco Unity, and others) as well as business application services (for example, the corporate email system and internal websites) within the enterprise network. With the CVO VPN connection established, Unified Communications soft clients and IP phones can register with Unified CM and make voice, video, and encrypted calls between enterprise devices.
- The Cisco VPN client for Cisco Unified IP Phones provides enterprise access for a subset of Cisco Unified IP Phone models. For more information on the devices supporting the Cisco VPN client for Cisco Unified IP Phones, see the chapter on Collaboration Endpoints, page 8-1. The phone's VPN client creates a tunnel (for the phone only), allowing to it register with Unified CM and to make voice, video, and encrypted calls between enterprise devices. A computer connected to the phone's PC port is responsible for authenticating and establishing its own tunnel to the enterprise with VPN client software.

VPN access gives the user access to all Unified Communications and business applications within the enterprise by creating a secure encrypted tunnel from the device to the VPN head-end. All traffic, including traffic destined for the internet and media for calls between VPN users, must always traverse the enterprise network rather than be established directly from the device over the internet to its destination. (See Figure 10-13.)

Figure 10-13 VPN Based Access



All of the above devices use their VPN client to connect to the enterprise network via a VPN head-end platform such as a Cisco Adaptive Security Appliance (ASA 5500) or a Cisco VPN aggregation router.

For more information on VPN access solutions, refer to the Unified Access and BYOD solutions guides available at

https://www.cisco.com/go/designzone

VPN-less Enterprise Access

Instead of using a VPN tunnel, VPN-less clients establish a secure and encrypted signaling path to an enterprise edge traversal platform such as Cisco Expressway. VPN-less clients register with Unified CM within the enterprise, and the secured channel to the edge traversal platform allows the client to establish an encrypted media path over the Internet for calls to other enterprise devices or to the PSTN through the enterprise PSTN gateway. Inside the enterprise signaling is typically unencrypted, whereas media can optionally remain encrypted.

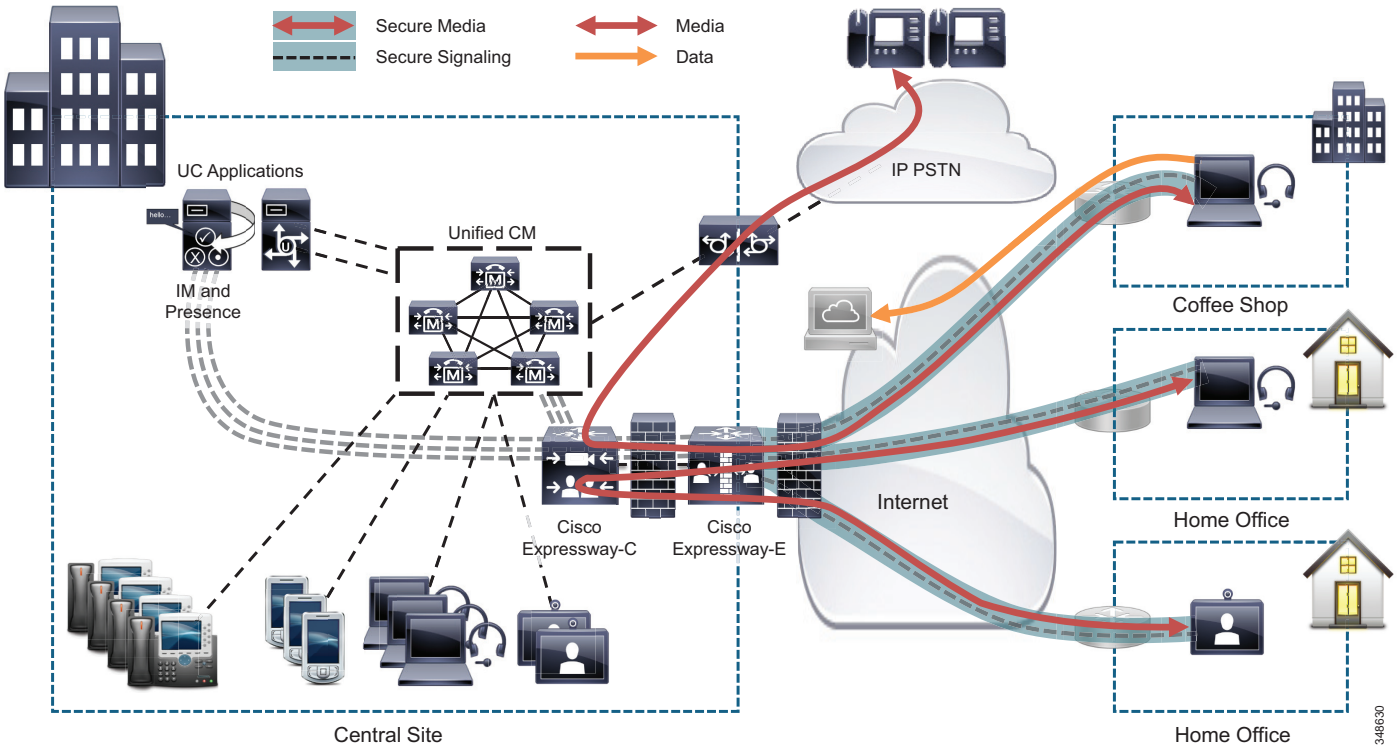
Unlike VPN clients, VPN-less clients provide enterprise access to Collaboration applications only; business applications within the enterprise (such as corporate email and internal websites) are not accessible, and connections to the Internet are made directly from the device rather than through the enterprise. Cisco VPN-less client access can be deployed using Cisco Expressway as the edge traversal platform.

This deployment type uses Cisco Expressway-C and Expressway-E. Cisco Expressway-E can be placed either in a DMZ or in the public internet, and it communicates by means of Cisco Expressway-C to the Unified CM cluster in the enterprise network. (See Figure 10-14.) Cisco Expressway supports VPN-less

access primarily for Cisco Jabber clients and TelePresence endpoints. Voice, video, encrypted calls, and IM and Presence are supported between enterprise endpoints. Media and signaling for calls between remote VPN-less devices traverse Cisco Expressway-C and Expressway-E. For specific information on the range of endpoints supported with Cisco Expressway VPN-less enterprise access, see the chapter on Collaboration Endpoints, page 8-1. For more information on Cisco Expressway VPN-less client access, refer to the documentation available at the following locations:

- https://www.cisco.com/c/en/us/solutions/collaboration/collaboration-edge-architecture/index.html
- https://www.cisco.com/c/en/us/products/unified-communications/expressway-series/index.html
- https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html

Figure 10-14 Collaboration Edge VPN-Less Access with Cisco Expressway



Business-to-Business Communications

Both Cisco Expressway and Cisco Unified Border Element (CUBE) support Internet based business-to-business unified communications connections between enterprises. Both Cisco Expressway and CUBE use SIP or H.323 trunks for business-to-business unified communications signaling. Cisco Expressway supports voice calls, video calls, and IM and Presence federation (see Figure 10-15); while CUBE supports voice calls and video calls only (see Figure 10-16).

Figure 10-15 Business-to-Business Communications Using Cisco Expressway

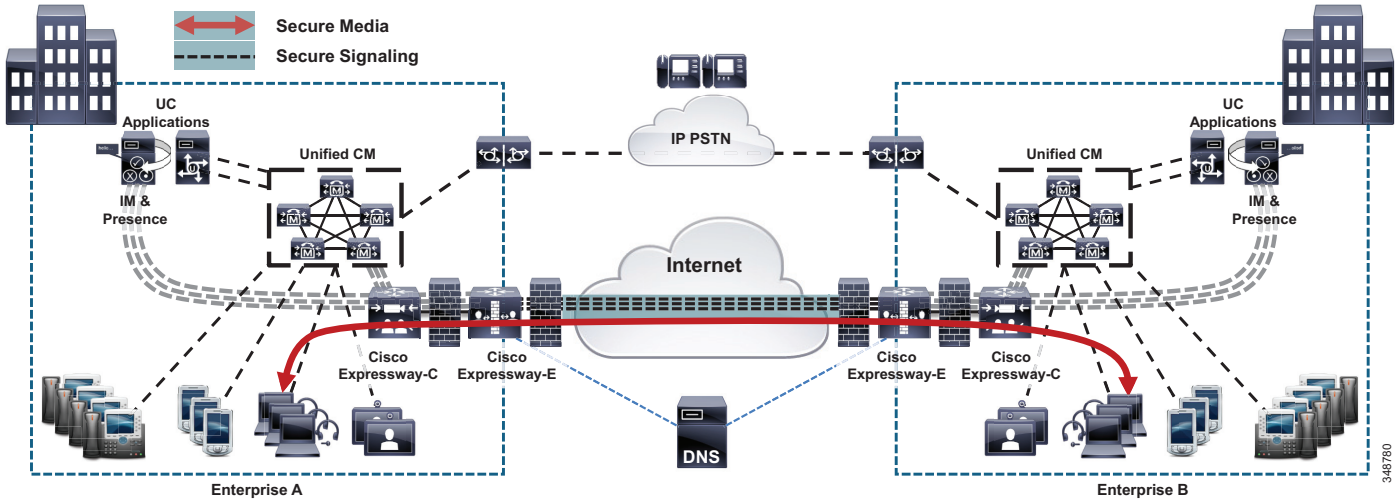
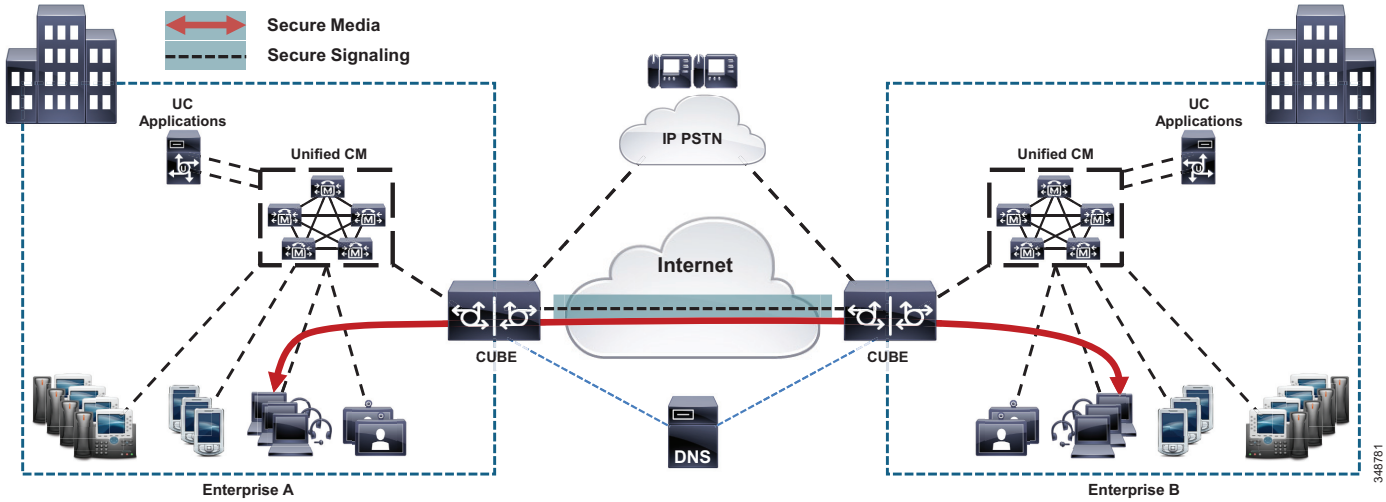


Figure 10-16 Business-to-Business Communications Using Cisco Unified Border Element (CUBE)



For additional information about deploying business-to-business communications with Cisco Expressway and Unified Border Element, refer to IP Gateways, page 5-15.

IP PSTN Deployments

IP PSTN deployments are increasing in popularity and are gradually replacing existing TDM-based PSTN access. SIP is commonly used as the IP PSTN access protocol, and today many service providers offer a voice-only service to the IP PSTN through a session border controller such a Cisco Unified Border Element. Session border controllers are SIP Back-to-Back User Agents (B2BUAs) and are typically used in flow-through mode, where both the voice media and SIP signalling for each call flow through Cisco Unified Border Element. (See Figure 10-17.) As a B2BUA in flow-through mode, Cisco

Unified Border Element can implement sophisticated QoS marking and call admission control policies while also providing support for transcoding, encryption, media forking for call recording applications, and scripting that allows SIP messages and SDP content to be modified for interoperability. For more info on Cisco Unified Border Element features and functions, refer to the latest version of the *Cisco Unified Border Element Data Sheet*, available at

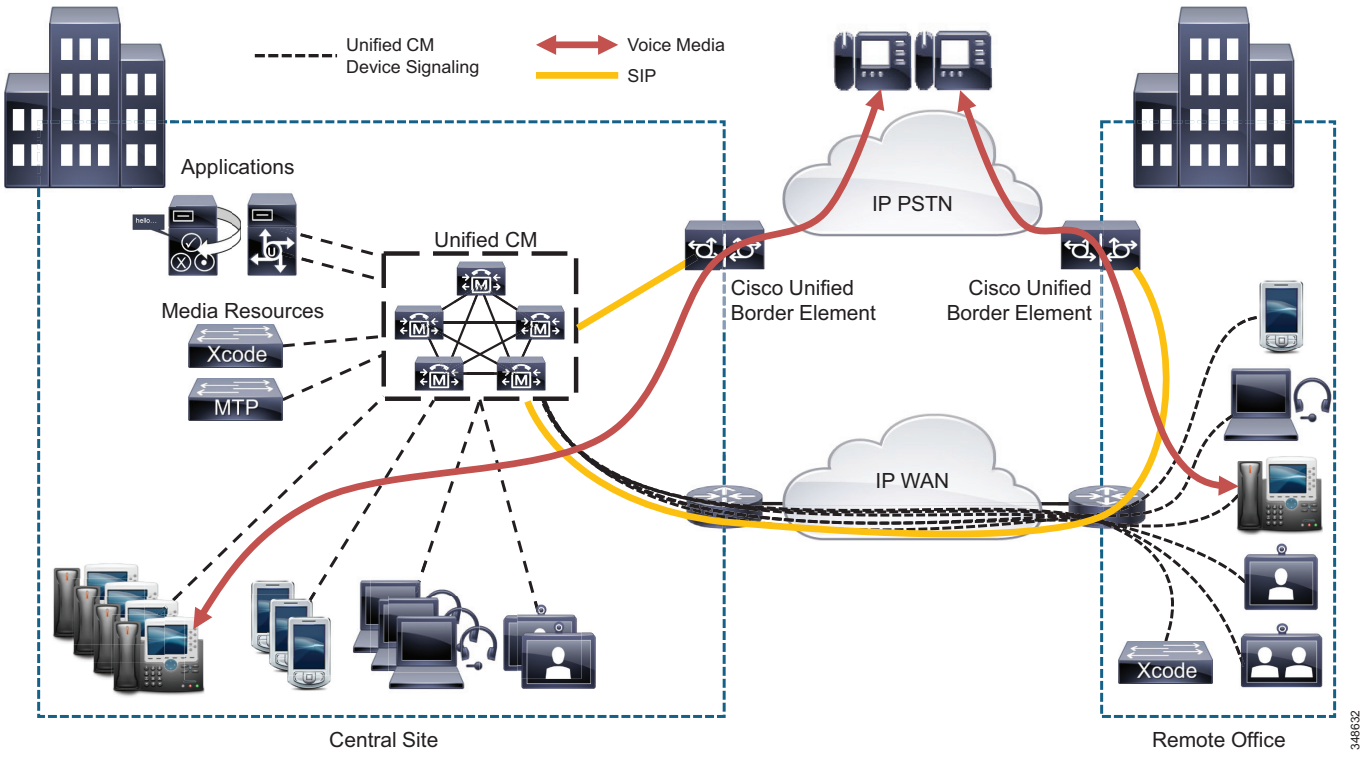
https://www.cisco.com/c/en/us/products/unified-communications/unified-border-element/index.html

Cisco Unified Border Element is supported on wide range of Cisco routing platforms, from the Cisco Integrated Services Routers (ISR) to the Cisco Aggregation Service Routers (ASR). Depending on the hardware platform, Cisco Unified Border Element can provide session scalability from 4 to 16,000 concurrent voice calls. Cisco Unified Border Element also provides redundancy on the following platforms:

- The Cisco ISR platforms, which can provide box-to-box redundancy with media preservation for stable active calls.
- The Cisco ASR platforms, which can provide box-to-box or in-box redundancy with media and signaling preservation (stateful failover) for stable active calls.

Note Access to the IP PSTN and access to the enterprise for VPN-less clients can be deployed on the same Cisco Unified Border Element platform.

Figure 10-17 Collaboration Edge IP PSTN Access



Geographic Deployment Options for IP PSTN

SIP trunks may be connected to IP PSTN service providers in several different ways, depending on the desired architecture. The two most common architectures for this connectivity are centralized trunks and distributed trunks.

Centralized trunks connect to the service provider (SP) through one logical connection (although there may be more than one physical connection for redundancy) with session border controllers (SBCs) such as the Cisco Unified Border Element. All IP PSTN calls to and from the enterprise use this set of trunks, and for most calls, media and signaling traverse the enterprise WAN to connect devices in the enterprise to those in the PSTN.

Distributed trunks connect to the service provider through several logical connections. Each branch of an enterprise may have its own local trunk to the service provider. With distributed trunks, media from the branch no longer needs to traverse the enterprise WAN but can flow directly to the service provider through a local SBC.

Each connectivity model has its own advantages and disadvantages. Centralized trunks are generally easier to deploy in terms of both physical equipment and configuration complexity. Distributed trunks have the advantage of local hand-off of media and better number portability from local providers. Alternatively, a hybrid connectivity model that combines some centralized and distributed IP PSTN access can capture the advantages of both forms of IP PSTN deployment.

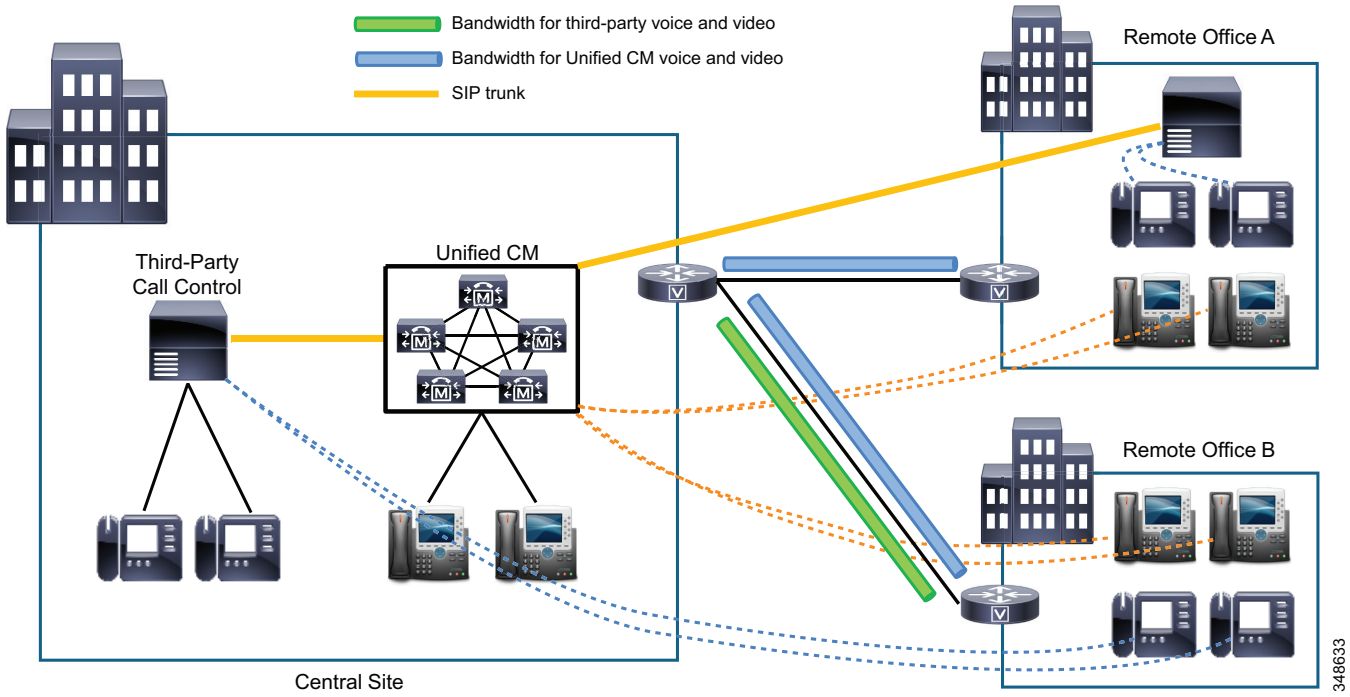
Design Considerations for Dual Call Control Deployments

In general, deployments where endpoints are registered to Cisco Unified CM and a third-party call control platform introduce a degree of complexity into any Collaboration Solution design, particularly with respect to dial plan and call admission control. In these dual call control deployments, both Unified CM and the third-party call control platform use independent mechanisms for call admission control and have independent dial plans, and the degree of complexity introduced is determined by the deployment model used. (See Figure 10-18.)

In campus deployments with dual call control, call admission control is not required and the dial plan in each call control system is relatively straight forward: If an endpoint cannot be found on one system, the call is forwarded to the other system. Standard dial plan configuration (calling search spaces and partitions) can be used to prevent routing loops between the systems.

In multisite centralized call processing deployments, a trade-off is typically made between call admission control complexity and dial plan complexity. If the Unified CM cluster and the third-party call control platform are deployed at the central site only, the dial plan is relatively straightforward, but the WAN bandwidth that each system uses for call admission control must be separately considered and provisioned for in the WAN. If additional third-party call control systems are deployed at remote sites where third-party endpoints reside, call admission control complexity can be avoided at the expense of a more fragmented dial plan. These trade-offs are discussed in more detail in the following sections.

Figure 10-18 Dual Call Control Deployments with Centralized and Distributed Third-Party Systems



Call Admission Control Considerations in Dual Call Control Deployments

Call admission control provides mechanisms for preventing the oversubscription of network bandwidth by limiting the number of calls that are allowed on the network at a given time based on overall call capacity of the call processing components and on network bandwidth.

In dual call control deployments, both Unified CM and the third-party call control platform use independent mechanisms for call admission control.

In multisite centralized call processing deployments, where the Unified CM cluster and the third-party call control platform are deployed at the central site only, consideration needs to be made in the WAN for the bandwidth that each platform uses for call admission control. At remote sites where third-party endpoints are registered to a locally deployed third-party call control system, these call admission control considerations can be avoided

Multisite Centralized Unified CM Deployments with Distributed Third-Party Call Control

Dual call control deployment models that use a third-party call control system at every site where third-party endpoints reside, can be tightly integrated with Unified CM call admission control. This may be achieved by using a SIP trunk to Unified CM from the third-party call control system at each site and configuring this SIP trunk in the same Unified CM location as the Unified CM endpoints residing at that site. Note, however, that while this approach resolves the call admission control issues for Unified CM and third-party call control, it does so at the expense of provisioning a large number of third-party call control systems, which in turn fragments the dial plan in the network.

Multisite Centralized Unified CM Deployments with Centralized Third-Party Call Control

Like Unified CM, a third-party call control system may be centralized to serve third-party endpoints residing at multiple sites over the WAN. With this type of deployment, the third-party call control system provides call admission control for calls between third-party endpoints at different sites; and likewise, the centralized Unified CM cluster provides call admission control for calls between Unified CM endpoints at different sites. Because the Unified CM cluster and the third-party call control system use independent call admission control mechanisms, at sites where both Unified CM and third-party endpoints reside in a centralized call control deployment, separate amounts of bandwidth must provisioned in the WAN for Unified CM call admission control and third-party call control. When calls are made between Unified CM endpoints and third-party endpoints in the same site, call admission control bandwidth will be decremented by both the third-party call control system and the Unified CM cluster even though the media path for the call might not traverse the WAN. Although not ideal from a call admission control perspective, this centralized call processing design is more cost effective in terms of hardware and it reduces dial plan fragmentation (because endpoints are registered to either the centralized Unified CM cluster or the centralized third-party call control system only).

A pragmatic approach should be taken for each dual call control deployment. From a strategic perspective, deploying a third-party call control system at every branch today might not make good commercial sense. However, if accurate call admission control is the design priority, then this deployment model might be appropriate. Likewise, for deployments with a centralized third-party call control and a centralized Unified CM cluster, the issue of independent call admission control domains can be addressed by over-provisioning bandwidth in the WAN.

Dial Plan Considerations in Dual Call Control Deployments

With only two call control systems in the deployment, if numbers are used to address endpoints and if the dialed number matches the local pattern of the local Unified CM cluster or third-party call control system, then the call will be sent to the locally attached endpoint. If the number matches the pattern for the other (remote) call control, then the call must be sent to the non-local endpoint through the interconnecting SIP trunk. In case of an overlapping dial plan, both the Unified CM cluster and the third-party call control system are able to send the call that does not match any internally registered endpoint, to the other call control cluster.

As the number of call control systems within a deployment increases, dial plan fragmentation also increases. This issue can be further exacerbated if endpoints from two or more call controls exist within the same site and do not have separate or easily summarized number ranges. In this case a default route cannot be used, but either of two options can be deployed to route calls in a Unified Communications system with multiple call control systems and a highly fragmented dial plan:

- Use an explicit route pattern and corresponding trunk for each of the unique number ranges associated with each call control.
- Within the Unified CM (and SME, if used) deployment, use the Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR) to share information about the number ranges supported by each Unified CM cluster and each third-party unified communications system. For third-party systems and their associated devices, import each unique number range into GDPR and associate each imported number range with a route string (a label that identifies the call control system). When a Unified CM user dials a number, Unified CM checks to see if the number is registered to its cluster. If the number is not registered to the Unified CM cluster, Unified CM searches ILS for the called number and its corresponding route string. The route string identifies the call control cluster where the number resides, which is used to match a SIP route pattern that then forwards the call over a SIP trunk toward its destination.

If alphanumeric URIs are used to address and call endpoints registered to Unified CM and the third-party call control system, then call routing can be implemented in either of the following ways, depending on the deployment:

- For deployments where only a single third-party call control system exists with a single SIP trunk to a Unified CM cluster, a default SIP route can be configured on Unified CM and the third-party call control system, so that calls to endpoints that are not found on one call control are sent to the other call control.
- If multiple third-party call control systems are deployed, use the Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR) to share information about the URIs supported by each Unified CM cluster and each third-party unified communications system. For URI-based call routing when a Unified CM user dials a URI, Unified CM checks to see if the URI is registered to its cluster. If it is not, Unified CM searches the ILS for the called URI and its corresponding route-string. The route string identifies the call control cluster where the URI resides, and it is used to match a SIP route pattern, which then forwards the call over a SIP trunk toward the destination URI. For URI-based endpoints registered to a third-party call control system, the list of URIs registered to the third-party call control system must be imported manually into ILS along with the corresponding route string for the third-party call control system.

Clustering Over the IP WAN

You may deploy a single Unified CM cluster (Enterprise Edition, Business Edition 7000, or Business Edition 6000) across multiple sites that are connected by an IP WAN with QoS features enabled. This section provides a brief overview of clustering over the WAN. For further information, refer to the chapter on Call Processing, page 9-1.

Clustering over the WAN can support two types of deployments:

- Local Failover Deployment Model, page 10-47

  Local failover requires that you place the Unified CM subscriber and backup servers at the same site, with no WAN between them. This type of deployment is ideal for two to four sites with Unified CM.

- Remote Failover Deployment Model, page 10-54

  Remote failover allows you to deploy primary and backup call processing servers split across the WAN. Using this type of deployment, you may have multiple sites with Unified CM subscribers being backed up by Unified CM subscribers at another site.

Note Remote failover deployments might require higher bandwidth because a large amount of intra-cluster traffic flows between the subscriber servers.

You can also use a combination of the two deployment models to satisfy specific site requirements. For example, two main sites may each have primary and backup subscribers, with another two sites containing only a primary server each and utilizing either shared backups or dedicated backups at the two main sites.

Some of the key advantages of clustering over the WAN are:

- Single point of administration for users for all sites within the cluster
- Feature transparency
- Shared line appearances

- Extension mobility within the cluster
- Unified dial plan

These features make this solution ideal as a disaster recovery plan for business continuance sites or as a single solution for multiple small or medium sites.

WAN Considerations

For clustering over the WAN to be successful, you must carefully plan, design, and implement various characteristics of the WAN itself. The Intra-Cluster Communication Signaling (ICCS) between Unified CM servers consists of many traffic types. The ICCS traffic types are classified as either priority or best-effort. Priority ICCS traffic is marked with IP Precedence 3 (DSCP 24 or PHB CS3). Best-effort ICCS traffic is marked with IP Precedence 0 (DSCP 0 or PHB BE). The various types of ICCS traffic are described in Intra-Cluster Communications, page 10-45, which also provides further guidelines for provisioning. The following design guidelines apply to the indicated WAN characteristics:

- Delay

  The maximum one-way delay between any two Unified CM servers should not exceed 40 ms, or 80 ms round-trip time. Measuring the delay is covered in Delay Testing, page 10-46. Propagation delay between two sites introduces 6 microseconds per kilometer without any other network delays being considered. This equates to a theoretical maximum distance of approximately 6,000 km for 40 ms delay or approximately 3,720 miles. These distances are provided only as relative guidelines and in reality will be shorter due to other delay incurred within the network.

- Jitter

  Jitter is the varying delay that packets incur through the network due to processing, queue, buffer, congestion, or path variation delay. Jitter for the IP Precedence 3 ICCS traffic must be minimized using Quality of Service (QoS) features.

- Packet loss and errors

  The network should be engineered to provide sufficient prioritized bandwidth for all ICCS traffic, especially the priority ICCS traffic. Standard QoS mechanisms must be implemented to avoid congestion and packet loss. If packets are lost due to line errors or other “real world” conditions, the ICCS packet will be retransmitted because it uses the TCP protocol for reliable transmission. The retransmission might result in a call being delayed during setup, disconnect (teardown), or other supplementary services during the call. Some packet loss conditions could result in a lost call, but this scenario should be no more likely than errors occurring on a T1 or E1, which affect calls via a trunk to the PSTN/ISDN.

- Bandwidth

  Provision the correct amount of bandwidth between each server for the expected call volume, type of devices, and number of devices. This bandwidth is in addition to any other bandwidth for other applications sharing the network, including voice and video traffic between the sites. The bandwidth provisioned must have QoS enabled to provide the prioritization and scheduling for the different classes of traffic. The general rule of thumb for bandwidth is to over-provision and under-subscribe.

- Quality of Service

  The network infrastructure relies on QoS engineering to provide consistent and predictable end-to-end levels of service for traffic. Neither QoS nor bandwidth alone is the solution; rather, QoS-enabled bandwidth must be engineered into the network infrastructure.

Intra-Cluster Communications

In general, intra-cluster communications means all traffic between servers. There is also a real-time protocol called Intra-Cluster Communication Signaling (ICCS), which provides the communications with the Cisco CallManager Service process that is at the heart of the call processing in each server or node within the cluster.

The intra-cluster traffic between the servers consists of the following:

- Database traffic from the IBM Informix Dynamic Server (IDS) database that provides the main configuration information. The IDS traffic may be re-prioritized in line with Cisco QoS recommendations to a higher priority data service (for example, IP Precedence 1 if required by the particular business needs). An example of this is extensive use of Extension Mobility, which relies on IDS database configuration.
- Firewall management traffic, which is used to authenticate the subscribers to the publisher to access the publisher's database. The management traffic flows between all servers in a cluster. The management traffic may be prioritized in line with Cisco QoS recommendations to a higher priority data service (for example, IP Precedence 1 if required by the particular business needs).
- ICCS real-time traffic, which consists of signaling, call admission control, and other information regarding calls as they are initiated and completed. ICCS uses a Transmission Control Protocol (TCP) connection between all servers that have the Cisco CallManager Service enabled. The connections are a full mesh between these servers. This traffic is priority ICCS traffic and is marked dependant on release and service parameter configuration.
- CTI Manager real-time traffic is used for CTI devices involved in calls or for controlling or monitoring other third-party devices on the Unified CM servers. This traffic is marked as priority ICCS traffic and exists between the Unified CM server with the CTI Manager and the Unified CM server with the CTI device.

Note For detailed information on various types of traffic between Unified CM servers, refer to the TCP and UDP port information in the latest version of the *System Configuration Guide for Cisco Unified Communications Manager*, available at https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html.

Unified CM Publisher

The publisher server replicates a partial read-only copy of the master database to all other servers in the cluster. Most of the database modifications are done on the publisher. If changes such as administration updates are made in the publisher's master database during a period when another server in the cluster is unreachable, the publisher will replicate the updated database when communications are re-established. Database modifications for user-facing call processing features are made on the subscriber servers to which the IP phones are registered. These features include:

- Call Forward All (CFA)
- Message Waiting Indication (MWI)
- Privacy Enable/Disable
- Do Not Disturb (DND) Enable/Disable
- Extension Mobility Login (EM)
- Monitor (for future use; currently no updates at the user level)

- Hunt Group Logout
- Device Mobility
- CTI Certificate Authority Proxy Function (CAPF) status for end users and application users
- Credential checking and authentication

Each subscriber replicates these changes to every other server in the cluster. Any other configuration changes cannot be made on the database during the period when the publisher is unreachable or offline. Most normal operations of the cluster, including the following, will *not* be affected during the period of publisher failure:

- Call processing
- Failover
- Registration of previously configured devices

Other services or applications might also be affected, and their ability to function without the publisher should be verified when deployed.

Call Detail Records (CDR) and Call Management Records (CMR)

Call detail records and call management records, when enabled, are collected by each subscriber and uploaded to the publisher periodically. During a period that the publisher is unreachable, the CDRs and CMRs are stored on the subscriber's local hard disk. When connectivity is re-established to the publisher, all outstanding CDRs are uploaded to the publisher, which stores the records in the CDR Analysis and Reporting (CAR) database.

Delay Testing

The maximum round-trip time (RTT) between any two servers must not exceed 80 ms. This time limit must include all delays in the transmission path between the two servers. Verifying the round trip delay using the **ping** utility on the Unified CM server will not provide an accurate result. The ping is sent as a best-effort tagged packet and is not transported using the same QoS-enabled path as the ICCS traffic. Therefore, Cisco recommends that you verify the delay by using the closest network device to the Unified CM servers, ideally the access switch to which the server is attached. Cisco IOS provides a extended ping capable to set the Layer 3 type of service (ToS) bits to make sure the ping packet is sent on the same QoS-enabled path that the ICCS traffic will traverse. The time recorded by the extended ping is the round-trip time (RTT), or the time it takes to traverse the communications path and return.

The following example shows a Cisco IOS extended ping with the IP Precedence set to 3 (ToS byte value set to 96):

```
Access_SW#ping
Protocol [ip]:
Target IP address: 10.10.10.10
Repeat count [5]:
Datagram size [100]:
Timeout in seconds [2]:
Extended commands [n]: y
Source address or interface:
Type of service [0]: 96
Set DF bit in IP header? [no]:
Validate reply data? [no]:
Data pattern [0xABCD]:
Loose, Strict, Record, Timestamp, Verbose[none]:
Sweep range of sizes [n]:
Type escape sequence to abort.
```

```
Sending 5, 100-byte ICMP Echos to 10.10.10.10, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/2/4 ms
```

Error Rate

The expected error rate should be zero. Any errors, dropped packets, or other impairments to the IP network can have an impact on the call processing performance of the cluster. This may be noticeable by delay in dial tone, slow key or display response on the IP phone, or delay from off-hook to connection of the voice path. Although Unified CM will tolerate random errors, they should be avoided to prevent impairing the performance of the cluster.

Troubleshooting

If the Unified CM subscribers in a cluster are experiencing impairment of intra-cluster communications due to higher than expected delay, errors, or dropped packets, some of the following symptoms might occur:

- IP phones, gateways, or other devices on a remote Unified CM server within the cluster might temporarily be unreachable.
- Calls might be disconnected or might fail during call setup.
- Users might experience longer than expected delays before hearing dial tone.
- Busy hour call completions (BHCC) might be low.
- The ICCS (SDL session) might be reset or disconnected.
- The time taken to upgrade a subscriber and synchronize its database with the publisher will increase.

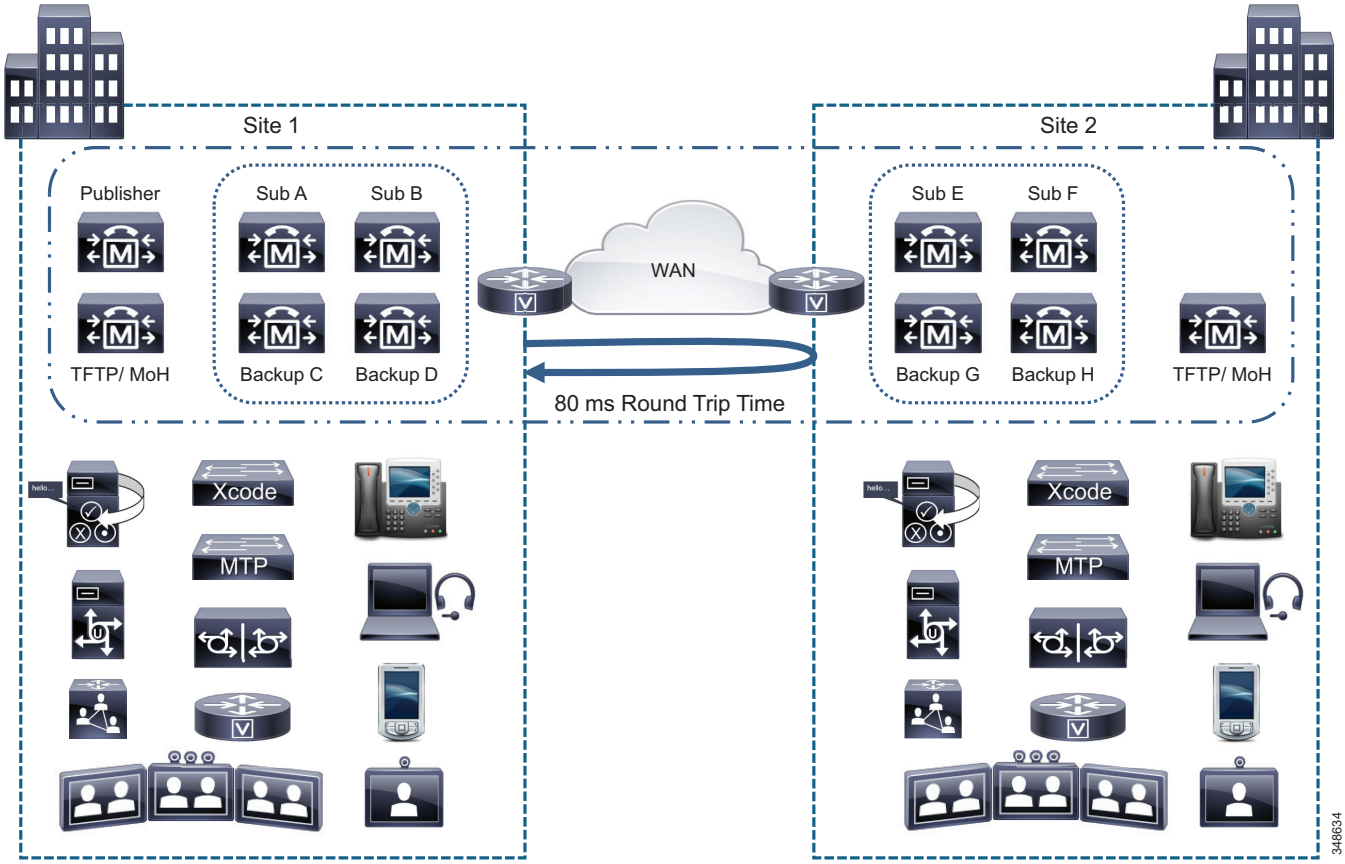
In summary, perform the following tasks to troubleshoot intra-cluster communication problems:

- Verify the delay between the servers.
- Check all links for errors or dropped packets.
- Verify that QoS is correctly configured.
- Verify that sufficient bandwidth is provisioned for the queues across the WAN to support all the traffic.

Local Failover Deployment Model

The local failover deployment model provides the most resilience for clustering over the WAN. Each of the sites in this model contains at least one primary Unified CM subscriber and one backup subscriber. This configuration can support up to four sites. The maximum number of phones and other devices will be dependent on the quantity and type of servers deployed. The maximum total number of IP phones for all sites is 40,000. (See Figure 10-19.)

Figure 10-19 Example of Local Failover Model



Observe the following guidelines when implementing the local failover model:

- Configure each site to contain at least one primary Unified CM subscriber and one backup subscriber.
- Configure Unified CM *groups* and *device pools* to allow devices within the site to register with only the servers at that site under all conditions.
- Cisco highly recommends that you replicate key services (TFTP, DNS, DHCP, LDAP, and IP Phone Services), all media resources (transcoders, conferencing resources, annunciator, and music on hold), and gateways at each site to provide the highest level of resiliency. You could also extend this practice to include a voicemail system at each site.
- Under a WAN failure condition, sites without access to the publisher database will lose some functionality. For example, system administration at the remote site will not be able to add, modify, or delete any part of the configuration. However, users can continue to access the user-facing features listed in the section on Unified CM Publisher, page 10-45.
- Under WAN failure conditions, calls made to phone numbers that are not currently communicating with the subscriber placing the call, will result in either a fast-busy tone or a call forward (possibly to voicemail or to a destination configured under Call Forward Unregistered).

- The maximum allowed round-trip time (RTT) between any two servers in the Unified CM cluster is 80 ms.

Note At a higher round-trip delay time and higher busy hour call attempts (BHCA), voice cut-through delay might be higher, causing initial voice clipping when a voice call is established.

- A minimum of 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) between each site and every other site that is clustered over the WAN. For example, if three sites are clustered over the WAN, each site would require 2 * 1.544 Mbps of WAN bandwidth for call control traffic. This minimum bandwidth requirement for call control traffic accounts for up to for 10,000 busy hour call attempts (BHCA) from one site to another site and applies only to deployments where directory numbers are not shared between sites that are clustered over the WAN. The following equation may be used as a guideline to calculate the bandwidth for more than 10,000 BHCA between non-shared directory numbers at a specific delay:

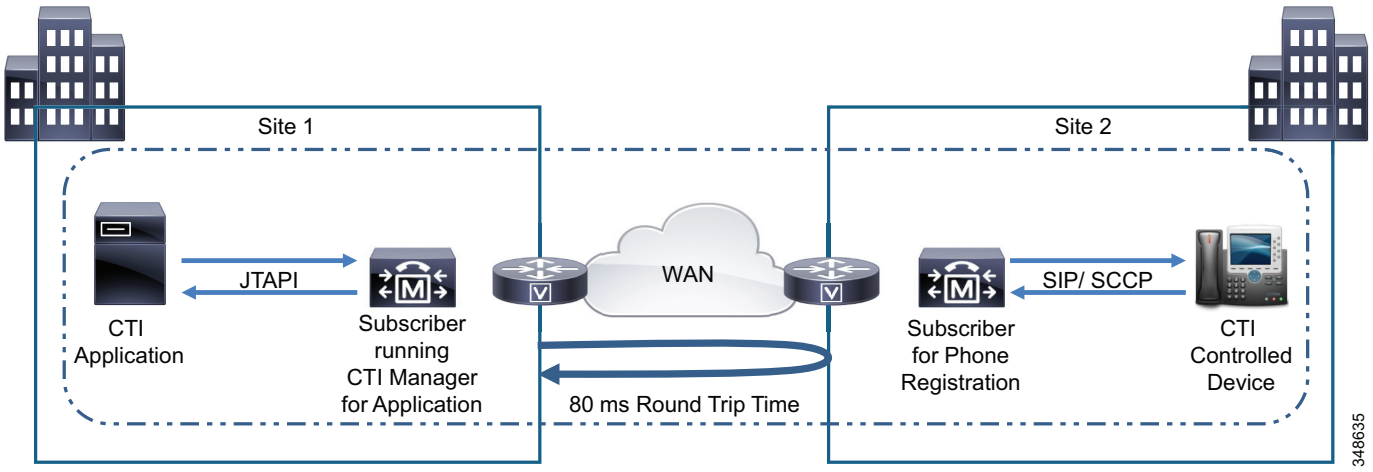
  Total Bandwidth (Mbps) = (Total BHCA/10,000) * (1 + 0.006 * Delay), where
  Delay = RTT delay in ms

  This call control traffic is classified as priority traffic. Priority ICCS traffic is marked with IP Precedence 3 (DSCP 24 or PHB CS3).

- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, a minimum of 1.544 Mbps (T1) bandwidth is required for database and other inter-server traffic between the publisher and every subscriber node within the cluster.
- For customers who also want to deploy CTI Manager over the WAN (see Figure 10-20), the following formula can be used to calculate the bandwidth (Mbps) for the CTI Intra-Cluster Communication Signaling (ICCS) traffic between the Unified CM subscriber running the CTI Manager service and the Unified CM subscriber to which the CTI controlled endpoint is registered:

  CTI ICCS bandwidth (Mbps) = (Total BHCA/10,000) * 0.53

Figure 10-20 CTI over the WAN

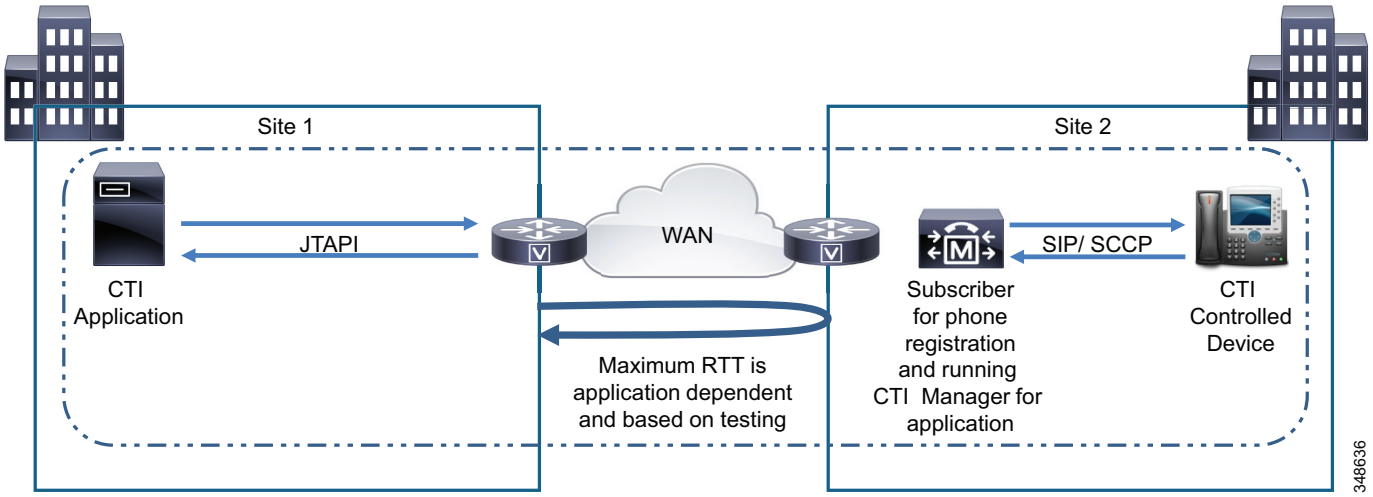


- For deployments where the J/TAPI application is remote from the Unified CM subscriber (see Figure 10-21), the following formula can be used to calculate the Quick Buffer Encoding (QBE) J/TAPI bandwidth for a typical J/TAPI application:

  J/TAPI bandwidth (Mbps) = (Total BHCA/10,000) * 0.28

  The bandwidth may vary depending on the J/TAPI application. Check with the application developer or provider to validate the bandwidth requirement.

Figure 10-21 J/TAPI over the WAN



Example 10-1 Bandwidth Calculation for Two Sites

Consider two sites, Site 1 and Site 2, with Unified CM clustered over the WAN across these two sites that are 80 ms round-trip time apart. Site 1 has one publisher, one combined TFTP and music on hold (MoH) server, and two Unified CM subscriber servers. Site 2 has one TFTP/MoH server and two Unified CM subscriber servers. Site 1 has 5000 phones, each having one DN; and Site 2 has 5000 phones, each having one DN. During the busy hour, 2500 phones in Site 1 call 2500 phones in Site 2, each at 3 BHCA. During that same busy hour, 2500 phones in Site 2 also call 2500 phones in Site 1, each at 3 BHCA. In this case:

Total BHCA during the busy hour = 2500*3 + 2500*3 = 15,000

Total bandwidth required between the sites = Total ICCS bandwidth + Total database bandwidth

Because total BHCA is 15,000 (greater than 10,000), we can use the formula to calculate:
Total ICCS bandwidth = (15,000/10,000) * (1 + 0.006*80) = 2.22 Mbps

Total database bandwidth = (Number of servers remote to the publisher) * 1.544 = 3 * 1.544 = 4.632 Mbps

Total bandwidth required between the sites = 2.22 Mbps + 4.632 Mbps = 6.852 Mbps (Approximately 7 Mbps)

- When directory numbers are shared between sites that are clustered over the WAN, additional bandwidth must be reserved. This overhead or additional bandwidth (in addition to the minimum 1.544 Mbps bandwidth) for 10,000 BHCA between shared DNs can be calculated using the following equation:

  Overhead = (0.012 * Delay * Shared-line) + (0.65 * Shared-line), where:

  Delay = RTT delay over the IP WAN, in ms

  Shared-line = Average number of additional phones on which a directory number is shared across the WAN.

  The following equation may be used as a guideline to calculate the bandwidth for more than 10,000 BHCA between shared directory numbers at a specific delay:

  Total bandwidth (Mbps) = (Total BHCA/10,000) * (1 + 0.006 * Delay + 0.012 * Delay * Shared-line + 0.65 * Shared-line), where:

  Delay = RTT delay in ms

  Shared-line = Average number of additional phones on which a directory number is shared across the WAN.

Example 10-2 Bandwidth Calculation for Two Sites with Shared Directory Numbers

Consider two sites, Site 1 and Site 2, with Unified CM clustered over the WAN across these two sites that are 80 ms round-trip time apart. Site 1 has one publisher, one combined TFTP and music on hold (MoH) server, and two Unified CM subscriber servers. Site 2 has one TFTP/MoH server and two Unified CM subscriber servers. Site 1 has 5000 phones, each having one DN; and Site 2 has 5000 phones, each sharing a DN with the 5000 phones in Site 1. Thus, each DN is shared across the WAN with an average of one additional phone. During the busy hour, 2500 phones in Site 1 call 2500 phones in Site 2, each at 3 BHCA. This also causes the phones in Site 1 to ring. During that same busy hour, 2500 phones in Site 2 call 2500 phones in Site 1, each at 3 BHCA. This also causes the phones in Site 2 to ring. In this case:

Total BHCA during the busy hour = 2500 * 3 + 2500 * 3 = 15,000

Total bandwidth required between the sites = Total ICCS bandwidth + Total database bandwidth

Because total BHCA is 15,000 (greater than 10,000), we can use the formula to calculate:
Total ICCS bandwidth = (15,000/10,000) * (1 + 0.006*80 + 0.012*80*1 + 0.65*1) = 4.635 Mbps

Total database bandwidth = (Number of servers remote to the publisher) * 1.544 = 3 * 1.544 = 4.632 Mbps

Total bandwidth required between the sites = 4.635 Mbps + 4.632 Mbps = 9.267 Mbps (Approximately 10 Mbps)

Note The bandwidth requirements stated above are strictly for ICCS, database, and other inter-server traffic. If calls are going over the IP WAN, additional bandwidth must be provisioned for media traffic, depending on the voice and video codecs used for the calls. For details see Bandwidth Provisioning, page 3-52.

- Subscriber servers in the cluster read their local database. Database modifications can occur in both the local database as well as the publisher database, depending on the type of changes. Informix Dynamic Server (IDS) database replication is used to synchronize the databases on the various servers in the cluster. Therefore, when recovering from failure conditions such as the loss of WAN connectivity for an extended period of time, the Unified CM databases must be synchronized with any changes that might have been made during the outage. This process happens automatically when

database connectivity is restored to the publisher and other servers in the cluster. This process can take longer over low bandwidth and/or higher delay links. In rare scenarios, manual reset or repair of the database replication between servers in the cluster might be required. This is performed by using the commands such as **utils dbreplication repair all** and/or **utils dbreplication reset all** at the command line interface (CLI). Repair or reset of database replication using the CLI on remote subscribers over the WAN causes all Unified CM databases in the cluster to be re-synchronized. With longer delays and lower bandwidth between the publisher and subscriber nodes, it can take longer for database replication repair or reset to complete.

Note Repairing or resetting of database replication on multiple subscribers at the same remote location can result in increased time for database replication to complete. Cisco recommends repairing or resetting of database replication on these remote subscribers one at a time. Repairing or resetting of database replication on subscribers at different remote locations may be performed simultaneously.

- If remote branches using centralized call processing with clustering over the WAN are connected to the central sites via the same WAN path that is used for clustering over the WAN traffic, pay careful attention to the configuration of call admission control to avoid oversubscribing the links used for clustering over the WAN.
  - If the bandwidth is not limited on the links used for clustering over the WAN (that is, if the interfaces to the links are OC-3s or STM-1s and there is no requirement for call admission control), then the remote sites may be connected to any of the main sites because all the main sites should be configured as location Hub_None. This configuration still maintains hub-and-spoke topology for purposes of call admission control.
  - If you are using the Multiprotocol Label Switching (MPLS) Virtual Private Network (VPN) feature, all sites in Unified CM locations and the remote sites may register with any of the main sites.
  - If bandwidth is limited between the main sites, call admission control must be used between sites, and all remote sites must register with the main site that is configured as location Hub_None. This main site is considered the hub site, and all other remote sites and clustering-over-the-WAN sites are spokes sites.
- During a software upgrade, all servers in the cluster should be upgraded during the same maintenance period, using the standard upgrade procedures outlined in the software release notes. The software upgrade time will increase for higher round-trip delay time over the IP WAN. Publisher to subscriber bandwidth lower than the required 1.544 Mbps for each subscriber node can also cause the software upgrade process to take longer to complete. If a faster upgrade time is desired, additional bandwidth above the required 1.544 Mbps per remote subscriber can be provisioned during the upgrade period.

Unified CM Provisioning for Local Failover

Provisioning of the Unified CM cluster for the local failover model should follow the design guidelines for capacities outlined in the chapter on Call Processing, page 9-1. If voice or video calls are allowed across the WAN between the sites, then you must configure Unified CM *locations* in addition to the default location for the other sites, to provide call admission control between the sites. If the bandwidth is over-provisioned for the number of devices, it is still best practice to configure call admission control based on locations. If the locations-based call admission control rejects a call, automatic failover to the PSTN can be provided by the automated alternate routing (AAR) feature.

To improve redundancy and upgrade times, Cisco recommends that you enable the Cisco Trivial File Transfer Protocol (TFTP) service on two Unified CM servers. More than two TFTP servers can be deployed in a cluster, however this configuration can result in an extended period for rebuilding all the TFTP files on all TFTP servers.

You can run the TFTP service on either a publisher or a subscriber server, depending on the site and the available capacity of the server. The TFTP server option must be correctly set in the DHCP servers at each site. If DHCP is not in use or if the TFTP server is manually configured, you should configure the correct TFTP address for the site.

Other services, which may affect normal operation of Unified CM during WAN outages, should also be replicated at all sites to ensure uninterrupted service. These services include DHCP servers, DNS servers, corporate directories, and IP phone services. On each DHCP server, set the DNS server address correctly for each location.

IP phones may have shared line appearances between the sites. During a WAN outage, call control for each line appearance is segmented, but call control returns to a single Unified CM server once the WAN is restored. During the WAN restoration period, there is additional traffic between the two sites. If this situation occurs during a period of high call volume, the shared lines might not operate as expected during that period. This situation should not last more than a few minutes, but if it is a concern, you can provision additional prioritized bandwidth to minimize the effects.

Gateways for Local Failover

Normally, gateways should be provided at all sites for access to the PSTN. The device pools should be configured to register the gateways with the Unified CM servers at the same site. Call routing (route patterns, route lists, and route groups) should also be configured to select the local gateways at the site as the first choice for PSTN access and the other site gateways as a second choice for overflow. Take special care to ensure emergency service access at each site.

You can centralize access to the PSTN gateways if access is not required during a WAN failure and if sufficient additional bandwidth is configured for the number of calls across the WAN. For E911 requirements, additional gateways might be needed at each site.

Voicemail for Local Failover

Cisco Unity Connection or other voicemail systems can be deployed at all sites and integrated into the Unified CM cluster. This configuration provides voicemail access even during a WAN failure and without using the PSTN. Using Voice Mail Profiles, you can allocate the correct voicemail system for the site to the IP phones in the same location. For more information on Unity Connection and clustering over then WAN, see Distributed Messaging with Clustering Over the WAN, page 19-16.

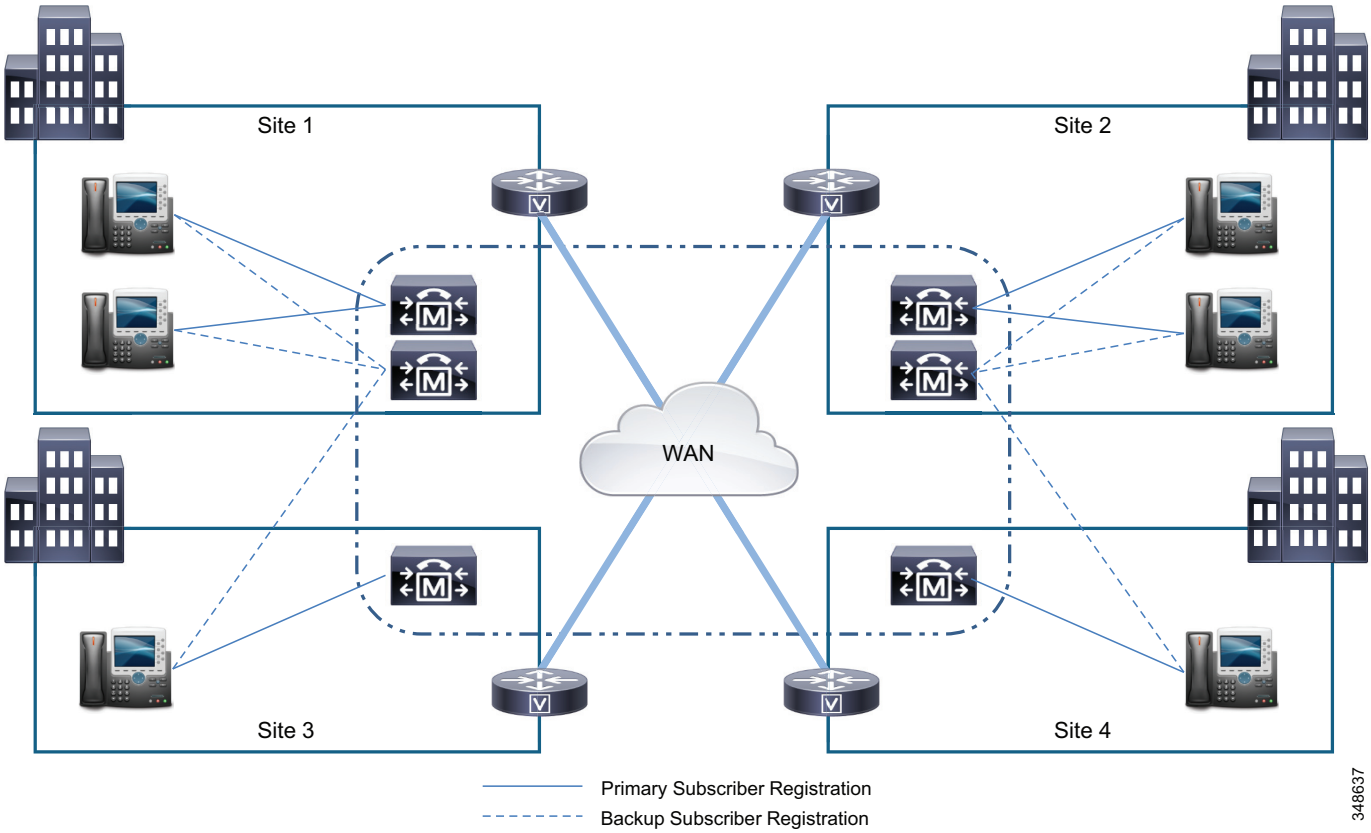
Music on Hold and Media Resources for Local Failover

Music on hold (MoH) servers and other media resources such as conference bridges should be provisioned at each site, with sufficient capacity for the type and number of users. Through the use of media resource groups (MRGs) and media resource group lists (MRGLs), media resources are provided by the on-site resource and are available during a WAN failure.

Remote Failover Deployment Model

The remote failover deployment model provides flexibility for the placement of backup servers. Each of the sites contains at least one primary Unified CM subscriber and may or may not have a backup subscriber. This model allows for multiple sites, with IP phones and other devices normally registered to a local subscriber when using 1:1 redundancy and the 50/50 load balancing option described in the chapter on Call Processing, page 9-1. Backup subscribers are located across the WAN at one or more of the other sites. (See Figure 10-22.)

Figure 10-22 Clustering over the WAN, Remote Failover Model with Four Sites



When implementing the remote failover model, observe all guidelines for the local failover model (see Local Failover Deployment Model, page 10-47), with the following modifications:

- Configure each site to contain at least one primary Unified CM subscriber and an optional backup subscriber as desired. If a backup subscriber over the IP WAN is not desired, a Survivable Remote Site Telephony (SRST) router may be used as a backup call processing agent.
- You may configure Unified CM *groups* and *device pools* to allow devices to register with servers over the WAN as a second or third choice.
- Signaling or call control traffic requires bandwidth when devices are registered across the WAN with a remote Unified CM server in the same cluster. This bandwidth might be more than the ICCS traffic and should be calculated using the bandwidth provisioning calculations for signaling, as described in Bandwidth Provisioning, page 3-52.

Note You can also combine the features of these two types of deployments for disaster recovery purposes. For example, Unified CM groups permit configuring up to three servers (primary, secondary and tertiary). Therefore, you can configure the Unified CM groups to have primary and secondary servers that are located at the same site and the tertiary server at a remote site over the WAN.

Deploying Unified Communications on Virtualized Servers

With virtualization, Cisco Collaboration application nodes are deployed as virtual machines (VMs) running on a physical server (host) via a hypervisor. Typically, multiple virtual machines can run on a host. This has obvious benefits over traditional deployments where the applications are directly running on the hardware platform. For example, costs (such as hardware, energy, cabling, and rack space costs) can be significantly reduced, and the operation and maintenance of the hardware platforms can be simplified by leveraging virtualization software capabilities.

This section presents a short introduction of the Cisco Unified Computing System (UCS) architecture, Hypervisor Technology for Application Virtualization, and Storage Area Networking (SAN) concepts. It also includes design considerations for deploying Unified Communications applications over virtualized servers.

This description is not meant to replace or supersede product-specific detailed design guidelines available at the following locations:

- https://www.cisco.com/c/en/us/products/servers-unified-computing/index.html
- https://www.cisco.com/go/virtualized-collaboration

For sizing aspects of Unified Communications systems on virtualized servers, use the Cisco Collaboration Sizing Tool, available to Cisco partners and employees (with valid login authentication) at

https://cucst.cloudapps.cisco.com/landing

Hypervisor

A hypervisor is a thin software system that runs directly on the server hardware to control the hardware, and it allows multiple operating systems (guests) to run on a server (host computer) concurrently. A guest operating system (such as that of Cisco Unified CM) runs on another level above the hypervisor. Hypervisors are one of the foundation elements in cloud computing and virtualization technologies, and they consolidate applications onto fewer servers.

Most of the Cisco Collaboration Systems applications are supported only with virtualization. This means that deploying the VMware vSphere ESXi hypervisor is required for those applications and that they cannot be installed directly on the server (bare metal).

VMware vCenter is a tool that helps to manage your virtual environment. With Tested Reference Configurations, VMware vCenter is not mandatory; however, it is strongly recommended when deploying a large number of hosts. With specification-based hardware, VMware vCenter is required.

Server Hardware Options

Two hardware options are available for deploying Cisco Collaboration applications with virtualization:

- Tested Reference Configurations (TRC), which are selected hardware configurations based on Cisco Unified Computing System (UCS) platforms. They are tested and documented for specific guaranteed performance, capacity, and application co-residency scenarios running "full-load" Cisco Collaboration System virtual machines.
- Specification-based hardware that provides more hardware flexibility and that, for example, adds support for other Cisco UCS and third-party servers that are listed in the *VMware Compatibility Guide* (available at (https://www.vmware.com/resources/compatibility/search.php).

Cisco Unified Computing System

Unified Computing is an architecture that integrates computing resources (CPU, memory, and I/O), IP networking, network-based storage, and virtualization, into a single highly available system. This level of integration provides economies of power and cooling, simplified server connectivity into the network, dynamic application instance repositioning between physical hosts, and pooled disk storage capacity.

The Cisco Unified Computing System is built from many components. But from a server standpoint, the UCS architecture is divided into the following two categories:

- Cisco UCS B-Series Blade Servers, page 10-56
- Cisco UCS C-Series Rack-Mount Servers, page 10-58

For more details on the Cisco Unified Computing System architecture, refer to the documentation available at

https://www.cisco.com/c/en/us/products/servers-unified-computing/index.html

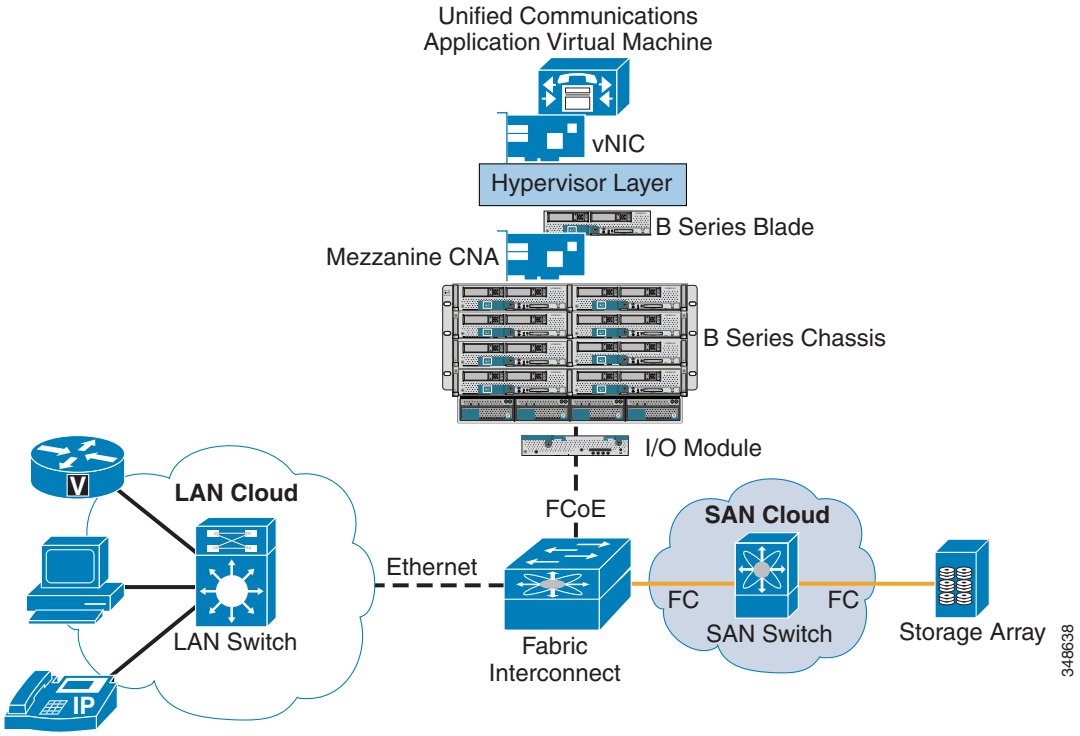
Cisco UCS E-Series server modules are blade servers designed to be deployed in Cisco Integrated Services Routers Generation 2 (ISR G2). Some Cisco Collaboration applications are supported on Cisco UCS E-Series, but the support might be limited (specification-based hardware support instead of TRC, for instance).

Cisco UCS B-Series Blade Servers

The Cisco Unified Computing System (UCS) features blade servers based on x86 architecture. Blade servers provide computing resources (memory, CPU, and I/O) to operating systems and applications. Blade servers have access to the unified fabric through mezzanine form factor Converged Network Adapters (CNA).

The architecture uses a unified fabric that provides transport for LAN, storage, and high-performance computing traffic over a single infrastructure with the help of technologies such as Fibre Channel over Ethernet (FCoE). (See Figure 10-23.) Cisco's unified fabric technology is built on a 10-Gbps Ethernet foundation that eliminates the need for multiple sets of adapters, cables, and switches for LANs, SANs, and high-performance computing networks.

Figure 10-23 Basic Architecture of Unified Communications on Cisco UCS B-Series Blade Servers



This section briefly describes the primary UCS components and how they function in a Unified Communications solution. For details about the Cisco UCS B-Series Blade Servers, refer to the model comparison at

https://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-b-series-blade-servers/models-comparison.html

Cisco UCS 5100 Series Blade Server Chassis

The Cisco UCS 5100 Series Blade Server chassis not only hosts the B-Series blade servers but also provides connectivity to the uplink Fabric Interconnect Switch by means of Cisco UCS Fabric Extenders.

Cisco UCS 2100 and 2200 Series I/O Modules

Cisco UCS 2100 and 2200 Series I/O Modules (or Fabric Extender) are inserted into the B-Series chassis, and they connect the Cisco UCS 5100 Series Blade Server Chassis to the Cisco UCS Fabric Interconnect Switch. The fabric extender can pass traffic between the blade server's FCoE-capable CNA to the fabric interconnect switch using Fibre Channel over Ethernet (FCoE) protocol.

Cisco UCS 6100 and 6200 Series Fabric Interconnect Switch

A Cisco UCS 6100 and 6200 Series Fabric Interconnect Switch is 10 Gigabit FCoE-capable switch. The B-Series Chassis (and the blade servers) connect to the fabric interconnect, and it connects to the LAN or SAN switching elements in the data center.

Cisco UCS Manager

Management is integrated into all the components of the system, enabling the entire UCS system to be managed as a single entity through the Cisco UCS Manager. Cisco UCS Manager provides an intuitive user interface to manage all system configuration operations.

Storage Area Networking

Storage area networking (SAN) enables attachment of remote storage devices or storage arrays to the servers so that storage appears to the operating system to be attached locally to the server. SAN storage can be shared between multiple servers.

Design Considerations for Running Virtual Unified Communications Applications on B-Series Blade Servers

This section highlights some design rules and considerations that must be followed for running Unified Communications services on virtualized servers.

Blade Server

The Cisco B-Series Blade Servers support multiple CPU sockets, and each CPU socket can host multiple multi-core processors. For example, one B200 blade has two CPU sockets that can host up to two multi-core processors. This provides the ability to run multiple Unified Communications applications on a single blade server. Each Unified Communications application should be allotted dedicated processing and memory resources to ensure that the resources are not oversubscribed.

SAN and Storage Arrays

Tested Reference Configurations based on the Cisco UCS B-Series platform require the virtual machines to run from a Fibre Channel SAN storage array. The SAN storage array must satisfy the requirements of the VMware hardware compatibility list. Other storage options such as iSCSI, FCoE SAN, and NFS NAS are supported with the specification-based hardware support. For more details, refer to the documentation available at

https://www.cisco.com/go/virtualized-collaboration

Cisco UCS C-Series Rack-Mount Servers

Beside the B-Series Blade Servers, the Cisco Unified Computing System (UCS) also features general purpose rack-mount servers based on x86 architecture. The C-Series Rack-Mount Servers provide computing resources (memory, CPU, and I/O) and local storage to the hypervisor and applications. For more information on C-Series servers, refer to the documentation at

https://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c-series-rack-servers/index.html

Design Considerations for Running Virtual Unified Communications Applications on C-Series Rack-Mount Servers

Unlike with UCS B-Series, the Tested Reference Configurations based on UCS C-Series support storage for the hypervisor and the applications virtual machines locally on the directly attached storage drives, not on an FC SAN storage array. It is possible to use an external storage array with a C-Series server, but the server would then be considered as specifications-based hardware and not as a TRC.

For more details, refer to the documentation available at

https://www.cisco.com/go/virtualized-collaboration

Impact of Virtual Servers on Deployment Models

Deploying Cisco Unified Communications applications on virtualized servers supports the same deployment models as when physical servers were used. There are a few additional considerations with virtualization, however. For example, the Unified CM VMware virtual application has no access to the host USB and serial ports. Therefore, Unified CM no longer supports the Cisco Messaging Interface (CMI) service for Simplified Message Desk Interface (SMDI) integrations, fixed MoH audio source integration for live MoH audio feeds using the audio cards (MOH-USB-AUDIO=), or flash drives to these servers. The following alternative options are available:

- For MoH live audio source feed, consider using Cisco IOS-based gateway multicast MoH for live audio source connectivity.
- For saving system install logs, use virtual floppy softmedia.

There is no alternative option for the Cisco Messaging Interface (CMI) service for Simplified Message Desk Interface (SMDI) integrations.

The chapter on Network Infrastructure, page 3-1, offers some design guidance on how to integrate the QoS capabilities of Cisco UCS B-Series virtualized servers into the network.

Call Routing and Dial Plan Distribution Using Call Control Discovery (CCD) for the Service Advertisement Framework (SAF)

The Cisco Service Advertisement Framework (SAF) is a Cisco IOS service routing protocol that can be used to share call routing and dial plan information automatically between call processing platforms. SAF allows non-Cisco call processing platforms (such as TDM PBXs) to partake in the Service Advertisement Framework when they are interconnected through a Cisco IOS gateway.

The Service Advertisement Framework (SAF) enables networking applications to advertise and discover information about networked services within an IP network. SAF consists of the following functional components and protocols:

- SAF Clients — Advertise and consume information about services.
- SAF Forwarders — Distribute and maintain SAF service availability information.
- The SAF Client Protocol — Used between SAF Clients and SAF Forwarders.
- The SAF Forwarder Protocol — Used between SAF Forwarders.

The nature of the advertised service is unimportant to the network of SAF Forwarders. The SAF Forwarder protocol is designed to dynamically distribute information about the availability of services to SAF client applications that have registered to the SAF network.

Services that SAF Can Advertise with Call Control Discovery (CCD)

In theory, any service can be advertised through SAF. The first service to use SAF is Cisco Unified Communications Call Control Discovery (CCD). CCD uses SAF to distribute and maintain information about the availability of internal directory numbers (DNs) hosted by call control agents such as Cisco Unified CM and Unified CME. CCD also distributes the corresponding number prefixes that allow these internal directory numbers to be reached from the PSTN ("To PSTN" prefixes).

Note SAF CCD supports the distribution of internal enterprise DN ranges only, unlike GDPR which supports the distribution of internal enterprise DN ranges, external (PSTN) DN ranges, and URIs.

The dynamic nature of SAF and the ability for call agents to advertise the availability of their hosted DN ranges and To PSTN prefixes to other call agents in a SAF network, provides distinct advantages over other static and more labor-intensive methods of dial plan distribution.

The following Cisco products support the Call Control Discovery (CCD) service for SAF:

- Cisco Unified Communications Manager (Unified CM)
- Cisco Unified Communications Manager Express (Unified CME) on a Cisco Integrated Services Router (ISR)
- Survivable Remote Site Telephony (SRST) on a Cisco ISR platform
- Cisco Unified Border Element on a Cisco ISR platform
- Cisco IOS Gateways on a Cisco ISR platform

CCD is supported on Cisco ISR platforms running Cisco IOS Release 15.0(1)M or higher.

For more information on SAF CCD in Unified Communications networks, refer to the SAF sections in the *Unified Communications Deployment Models* chapter of the *Cisco Unified Communications System 9.0 SRND*, available at

https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cucm/srnd/9x/uc9x/models.html

For more information on SAF itself, refer to the *Service Advertisement Framework (SAF)* section in the *Network Infrastructure* chapter of the *Cisco Collaboration 9.x Solution Reference Network Designs (SRND)*, available at

https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cucm/srnd/collab09/clb09/netstruc.html

SAF CCD Deployment Considerations

The following scalability limits apply to Unified CM and Cisco IOS SAF CCD products:

- Up to 2,000 advertised DN patterns per Unified CM cluster
- Up to 100,000 learned DN patterns per Unified CM cluster (Default value = 20,000 learned patterns)
- Up to 125 advertised DN patterns per Unified CME, Cisco Unified Border Element, or Cisco IOS Gateway
- Up to 6,000 learned DN patterns per Unified CME, Cisco Unified Border Element, Cisco IOS Gateway, or SRST (platform-dependant)

Note For SAF deployments using a single SAF autonomous system (AS) and consisting of Cisco Unified CM and SAF CCD running on a Cisco IOS platform, SAF CCD system-wide scalability is limited to 6,000 learned DN patterns.

Cisco Rich Media Conferencing

Revised: March 1, 2018

Conferencing is an essential component of any collaboration system, especially when serving remote users and/or a large user base. Cisco Rich Media Conferencing offers features such as instant, permanent, and scheduled audio and video conferencing, as well as content sharing.

Conference bridges provide the conferencing function. A conference bridge is a resource that joins multiple participants into a single call (audio or video). It can accept any number of connections for a given conference, up to the maximum capacity allowed for a single conference on that device. The output display for a given party shows all connected parties minus the viewer's own input.

Cisco Rich Media Conferencing solutions utilize various infrastructures to provide audio and video conferencing capability and content sharing. The conferencing infrastructure can be Cisco Unified CM using software or DSP resources, Cisco Meeting Server, or Cisco WebEx Collaboration Cloud, and this chapter covers the design details pertaining to each solution.

Cisco Rich Media Conferencing solutions are available as on-premises, cloud, or hybrid deployments. This allows an organization to integrate with the Collaboration solution in which they have already invested or, alternatively, to implement a service that is hosted "in the cloud." This is one of the more important distinctions between the various solutions, and it is the first decision point when determining which solution is the best fit for an organization.

Cisco WebEx Software as a Service (SaaS) offers a completely off-premises solution, while Cisco Collaboration Meeting Rooms (CMR) Hybrid is a hybrid solution with a mix of on-premises and off-premises equipment. Organizations that have deployed Cisco Collaboration System will benefit most from leveraging an on-premises solution. The later sections of this chapter provide more detailed deployment options for each conferencing solution.

Table 11-1 summarizes available solutions from an on-premises cloud perspective.

Table 11-1 On-Premises, Cloud, and Hybrid Capabilities of Cisco Collaborative Solutions

Solution	Audio		Video		Content Sharing	
	On-premises	Cloud	On-premises	Cloud	On-premises	Cloud
Cisco WebEx Meetings Server	Yes	No	Yes[1]	No	Yes	No
Cisco WebEx SaaS	No	Yes	No	Yes 1	No	Yes
Cisco Meeting Server	Yes	No	Yes	No	Yes	No
Cisco CMR Hybrid	Yes	Yes	Yes	Yes	Yes	Yes
Cisco WebEx Meeting Center Video Conferencing	No	Yes	No	Yes	No	Yes

1. Cisco WebEx webcam video only, and no support with standards-based video.

To provide a satisfactory end-user experience, careful planning and design should be done when deploying Cisco Conferencing solutions so that users are enabled with the conferencing functionality they require.

To aid in the design, this chapter starts with an introduction of the different types of conferences supported in the Cisco Conferencing solutions, followed by detailed discussions of the following main topics for each solution:

- Architecture

  This section introduces the main components of the Conferencing solution and describes its advantages as well as the different conferencing mechanisms available through the various components of a collaboration system. Supported deployment models, solutions, and recommendations are discussed here as well.

- High availability

  This section discusses best practices for designing a resilient Cisco Conferencing solution; it also contains guidance for redundancy and load balancing.

- Capacity planning

  This section provides best practices and design information related to capacity limits and scalability for the Cisco Conferencing solution.

- Design considerations

  This section discusses general recommendations and best practices for the Cisco Conferencing solution design.

This chapter contains discussions on the following Cisco Conferencing solutions:

- Cisco WebEx Software as a Service, (SaaS)
- Cisco WebEx Meetings Server – for private cloud
- Cisco WebEx Meeting Center Video Conferencing
- Cisco Meeting Server
- Cisco Collaboration Meeting Rooms (CMR) Hybrid

What's New in This Chapter

This chapter has been updated with new support and updated designs for Cisco Collaboration System Release (CSR) 12.*x*. You should read this entire chapter before deploying Conferencing in your Cisco Collaboration System.

Table 11-2 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 11-2 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Cisco Collaboration Meeting Rooms (CMR) Premises has been replaced by Cisco Meeting Server	Cisco Meeting Server, page 11-7	March 1, 2018
Content reorganization and other updates for Cisco Collaboration System Release (CSR) 12.*x*	Various sections of this chapter	March 1, 2018

Types of Conferences

The Cisco Rich Media Conferencing solution supports the following types of conferences:

- Instant conference

  An instant audio or video conference (also referred to as an ad hoc conference) is an impromptu conference. Instant conferences are not scheduled or arranged prior to the conference. For example, a point-to-point call escalated to a multipoint conference is considered to be an instant conference.

- Permanent conference

  Permanent conferences (also referred to as meet-me, static, or rendezvous conferences) are predefined addresses that allow conferencing without previous scheduling. The conference host shares the address with other users, who can call in to that address at any time.

  Permanent conference resources are used on a first-come-first-served basis (non-assured). For a guaranteed conference resource (assured), scheduled conferences should be used.

- Scheduled conference

  A scheduled conference is started by its initiator through a scheduling management system called Cisco TelePresence Management Suite (TMS). Conferences are booked via Cisco TMS with a start and end time and optionally with a predefined set of participants.

Cisco Rich Media Conferencing consists of the conferencing solutions described below. The details pertaining to each solution are described in each individual section that follows.

- Cisco Unified CM Audio Conferencing, page 11-4

  This solution allows Unified CM to use its internal software component or external hardware digital signal processors (DSPs) as the resources to perform audio conferencing.

- Cisco Meeting Server, page 11-7

  Cisco Meeting Server is an on-premises video conferencing solution. Each user has a personal Space that can be used to conduct meetings. Users can manage items such Space creation, adding members to a Space, and PIN creation from the Cisco Meeting App.

- Cisco Collaboration Meeting Rooms Hybrid, page 11-49

  Cisco CMR Hybrid combines the on-premises video conference and the WebEx Meeting Center conference into a single meeting, which allows TelePresence and WebEx participants to join and share voice, video, and content. CMR Hybrid meetings can be either scheduled or non-scheduled.

- Cisco WebEx Meeting Center Video Conferencing, page 11-34

  Cisco WebEx Meeting Center Video Conferencing (formerly Cisco Collaboration Meeting Rooms (CMR) Cloud) is an alternate conferencing deployment model that does not require any on-premises conferencing resources or management infrastructure. It supports both scheduled and non-scheduled meetings as well as TelePresence, audio, and WebEx participants in a single call, all hosted in the cloud.

- Cisco WebEx Meetings Server, page 11-41

  Where cloud-based web and audio conferencing is not suitable, it is possible to use the on-premises WebEx Meetings Server solution. This product offers a standalone audio, video, and collaboration web conferencing platform.

Cisco Unified CM Audio Conferencing

Cisco Unified CM supports audio conferences using any of the following methods:

- Software Audio Conferencing, page 11-4
- Hardware Audio Conferencing, page 11-5
- Built-in Bridge, page 11-5
- Cisco Conference Now, page 11-5

Software Audio Conferencing

The software-based audio conference bridges are provided by the IP Voice Media Streaming Application on Unified CM. The application must be enabled on each individual node in a cluster. A software unicast conference bridge is a standard conference mixer that is capable of mixing G.711 audio streams and Cisco Wideband audio streams. Any combination of Wideband or G.711 a-law and mu-law streams may be connected to the same conference. The number of conferences that can be supported on a given configuration depends on the server where the conference bridge software is running and on what other functionality has been enabled for the application. However, 256 is the maximum number of audio streams for this type. With 256 streams, a software conference media resource can handle 256 users in a single conference, or the software conference media resource can handle up to 64 conferencing resources with four users per conference. If the Cisco IP Voice Media Streaming Application service runs on the same server as the Cisco CallManager Service, a software conference should not exceed the maximum limit of 48 participants.

The Cisco IP Voice Media Streaming Application is a resource that can also be used for several functions, and the design must consider all functions together (see Cisco IP Voice Media Streaming Application, page 7-3). Since the capabilities of the software audio conference bridge are limited, Cisco recommends using a software audio conference bridge only in centralized deployments or in deployments where the use of a G.711 codec is acceptable for instant and meet-me audio conferencing. It is also important to note that the use of a software audio conference bridge in Unified CM will result in a higher load on the system than otherwise would be present.

Hardware Audio Conferencing

Digital signal processors (DSPs) that are configured through Cisco IOS as conference resources load firmware into the DSPs that are specific to conferencing functionality only, and these DSPs cannot be used for any other media feature. Any Cisco PVDM hardware may be used simultaneously in a single chassis for voice termination but may not be used simultaneously for other media resource functionality. DSPs on PVDM hardware are configured individually as voice termination, conferencing, media termination, or transcoding, so that DSPs on a single PVDM may be used as different resource types. Allocate DSPs to voice termination first, then to other functionality as needed. The DSP resources for a conference are reserved during configuration, based on the profile attributes and irrespective of how many participants actually join.

Hardware audio conference bridges offer a wider range of capabilities and codec format support than the software conference bridges. Cisco recommends using hardware audio conference bridges where the enterprise requires a more versatile audio conference bridge and codec support for higher-complexity codecs such as G.729 to take advantage of bandwidth savings.

Built-in Bridge

Built-in bridge refer to the DSP resources that are hosted by one of the endpoints in the call. Certain Cisco IP Phones have an on-board DSP for the built-in bridge functionality. The IP phone built-in bridge is the only embedded audio resource in the Cisco Rich Media Conferencing architecture. The built-in bridge, however, has limited conference functionality and cannot be used to launch a full conference. The built-in bridge in the Cisco IP Phones allows a user to:

- Join calls across different lines that the IP phone might have, and convert those calls into a conference hosted on the built-in bridge.
- Barge into a call of a different endpoint that shares the line (if the call is not set to private), and convert the call into a conference hosted on the built-in bridge.
- Start a silent recording or monitoring session from the endpoint that is engaged on a call, and fork the media generated and received by the phone invoking the feature.

The built-in bridge of the Cisco IP Phones can encode and decode G.711 and G.729 codec formats. However, once the codec for the call has been selected, the built-in bridge codec selection is locked and the phone will be unable to change the codec used. Therefore, the best practice is to carefully analyze the call flow in which the built-in bridge might be invoked to avoid call drops.

The built-in bridge can mix a maximum of two calls and can fork only one call (two streams).

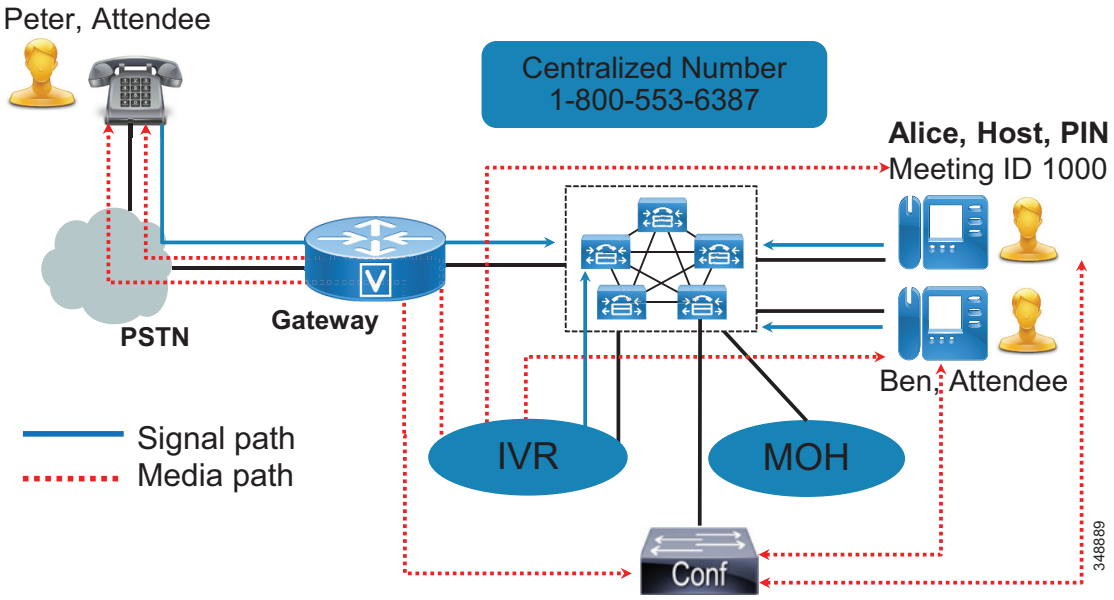
Cisco Conference Now

Cisco Conference Now is a Unified CM native application that provides permanent conference capability similar to Meet-Me. This application is targeted for small business customers who require a basic audio conferencing solution. With Cisco Conference Now, user can simply call into a centralized number and enter the appropriate meeting ID and a host or attendee PIN when prompted by the voice-guided system to join the conference.

Figure 11-1 shows the Cisco Conference Now architecture along with the components involved. Conference Now allows both external and internal callers to join a conference by dialing the Conference Now IVR directory number, which is a centralized conference assistant number. An IVR device is used

to guide and collect information from the caller to join the conference by playing announcements. The IVR is a media resource device that enables Unified CM to play prerecorded announcements (.wav files) to devices such as Cisco IP Phones and gateways.

Figure 11-1 Cisco Conference Now Architecture



The administrator can enable the Conference Now option for a user. If enabled, the user gets a meeting number and must configure a host PIN to start the meeting. Also, an optional attendee access code can be configured for attendees to join the meeting. Prior to the meeting, the conference host distributes the meeting number and the optional access code to all the participants. To start the meeting, the host dials into Conference Now and enters both the meeting number and the host PIN. To join the meeting, the attendees dial into Conference Now and enter the meeting number along with the optional attendee access code. If the attendee dials into the meeting before the host, the attendee will be placed on Music on Hold (MoH). Conference Now uses conference bridges configured in the media resource group (MRG) and media resource group list (MRGL) associated with the host's calling device to perform the conferencing function. Ensure that both the conference bridge and the IVR resources are available to Unified CM in order to use the Conference Now feature.

Using a conference bridge other than the software-based Cisco IP Voice Media Streaming Application (IPVMS) from Unified CM might not provide the conference party entry and exit tone. For the best user experience, we recommend using the software-based Cisco IPVMS conference bridge for Conference Now. For detail on conference party entry and exit tone support, refer to the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

Consider the following points when implementing Cisco Conference Now:

- The IVR supports Out of Band DTMF only. Use an MTP to convert any DTMF capability mismatch.
- The IVR supports G.711 (a-law and mu-law), G.729, and Wide Band 256K. IPVMS supports G.711 and Wide Band 256K. For other codec support, use a transcoder.
- Conference Now does not support any advanced functionality such as a roster list or muting and un-muting attendees.

Cisco Meeting Server

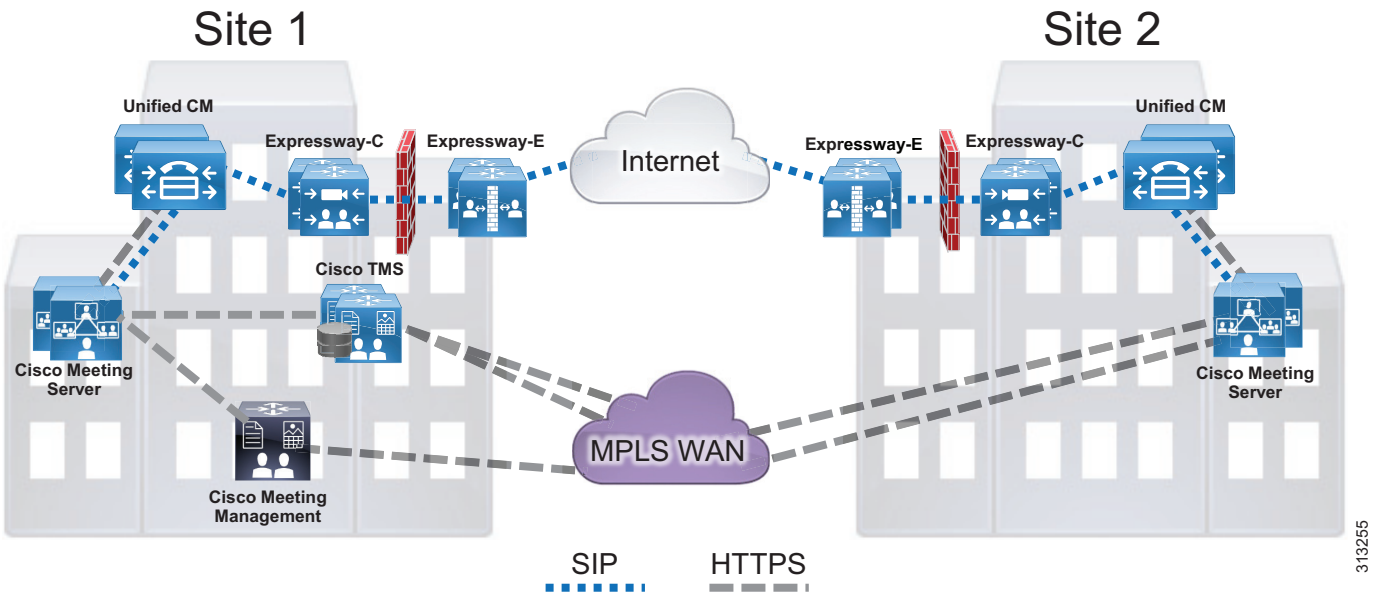
Cisco Meeting Server utilizes the Cisco on-premises infrastructure to provide business quality video and audio conferences as well as content sharing. Each user in the system can have an always-on personal Space with an associated video address (DN and/or URI) for participants to dial in and join the meeting. Cisco Meeting Server enables rich conferencing collaboration capabilities for endpoints registered to Cisco Unified Communications Manager (Unified CM) or Cisco Expressway as well as the ability to integrate business-to-business audio and video systems and legacy H.323 video systems interworked via Cisco Expressway. This architecture provides a rich feature set by relying on a variety of components in the conferencing solution. The following sections present an overview of those components and their roles in the Cisco Meeting Server conferencing solution.

Architecture

Figure 11-2 illustrates the architecture of the conferencing solution using Cisco Meeting Server. Cisco Meeting Server provides and manages conference resources; Cisco TelePresence Management Suite (TMS) provides the provisioning and scheduling functions for the conference resources; Cisco Meeting Management provides the meeting control and management functions; and Cisco Unified Communications Manager (Unified CM) is the call control system. Alternatively, Cisco Expressway can be used in place of Cisco Unified CM as the call control system. Cisco Meeting Server supports SIP call control only but can use Cisco Expressway to interwork with legacy H.323 video systems. Cisco Meeting Server is the conference bridge that supports all conference types, and it connects with Unified CM via a SIP trunk.

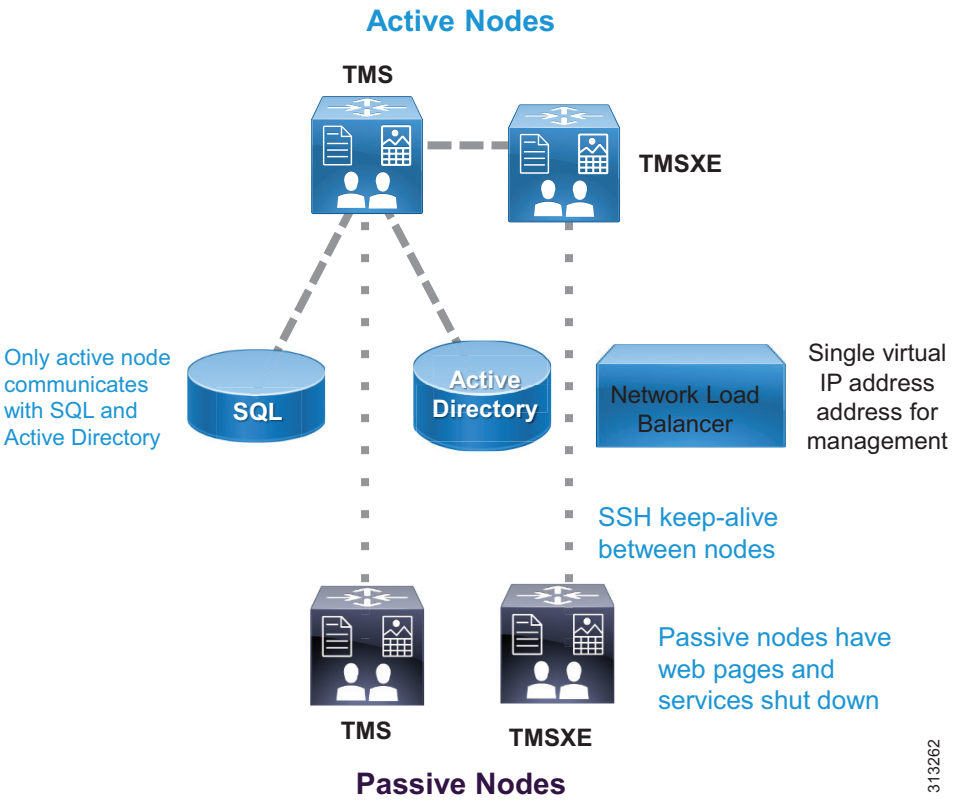
Cisco Unified CM communicates with Cisco Meeting Server using XML-RPC over HTTPS to control the conference bridges for instant conferences. Cisco TMS uses the REST API connections to link the Cisco Meeting Server to provision and schedule conference resources. Cisco Meeting Management has two interfaces with Cisco Meeting Server: REST API and Call Detail Record (CDR). The REST API interface is used to perform operations on Cisco Meeting Server, while the CDR interface is used to receive call events from Cisco Meeting Server.

Figure 11-2 Cisco Meeting Server Architecture Overview



The scheduling architecture consists of an active and a passive node for both Cisco TMS and TelePresence Management Suite Extension for Microsoft Exchange (TMSXE), which are deployed behind a network load balancer. The active node processes the incoming requests, while the passive node runs in standby mode with its web pages and services locked down and refusing all incoming traffic. For large deployments, Cisco TMS and TMSXE must be installed on separate virtual machines, as indicated in Figure 11-3. Cisco TMS servers are installed in the customer data center that also hosts the organization's SQL deployment. All the server nodes function from an external Microsoft SQL database. Additionally, endpoints, Cisco Meeting Server, and Unified CM make up the complete conference components for scheduling.

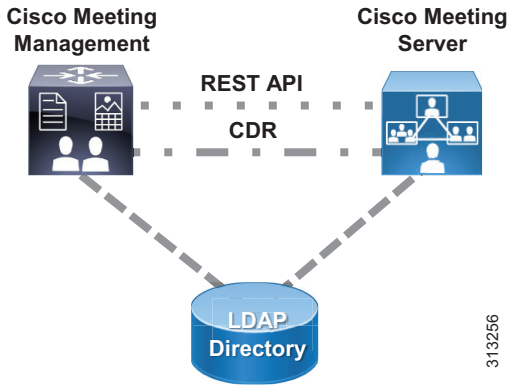
Figure 11-3 High-Level View of the Scheduling Architecture



Cisco Meeting Management runs on a separate server outside of Cisco Meeting Server and is dedicated to the Cisco Meeting Server deployment only. As mentioned previously, Cisco Meeting Management uses the REST API link to perform operations on Cisco Meeting Server. Cisco Meeting Management uses the CDR interface to receive call-related events from Cisco Meeting Server so that it knows when a meeting has started or ended, along with other call activities. When users log into the portal, they are authenticated against the LDAP directory. In addition, Cisco Meeting Management uses the groups configured inside the directory to determine the user's role (video operator or administrator). Based on the user's role, different options will be available on the Cisco Meeting Management portal. (See Figure 11-4.)

Note Cisco Meeting Server 2.1.5 is the minimum version required to deploy Cisco Meeting Management, but version 2.2 or later is recommended.

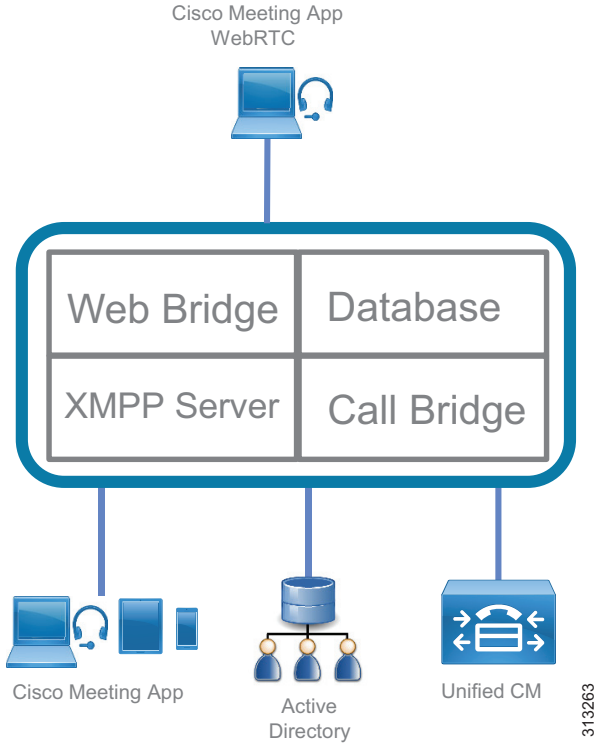
Figure 11-4 Cisco Meeting Management Architecture



Role of Cisco Meeting Server

Figure 11-5 illustrates the core conferencing components of Cisco Meeting Server that provide video conferencing capability. The call bridge component integrates with Cisco Unified CM for call control and provides resources to perform conference functions. All Cisco Meeting Server conferences are hosted on the Spaces. Spaces are virtual meeting rooms that have audio, video, and content sharing capability and are accessible using the Space URI or directory number. Cisco Meeting Server must integrate with a directory server such as Microsoft Active Directory to import users into the system. During the import process, Spaces are created using the field mapping expressions configuration. All the information for users and Spaces is stored in the database. Participants can join conferences using Cisco or third-party standard SIP video endpoints, Cisco Jabber clients, or the Cisco Meeting App. The XMPP server authenticates users logging in through the Cisco Meeting App. The Web Bridge connects WebRTC client users to the call bridge after they log in.

Figure 11-5 Core Conferencing Components of Cisco Meeting Server



Cisco Meeting App is the client to Cisco Meeting Server, and it can be a native desktop or mobile application or a WebRTC browser application. With Cisco Meeting App, users can log in and join the conference with audio and video along with content sharing. With the WebRTC client, users without an account on Cisco Meeting Server can use a compatible browser to join the conference as a guest. In addition, users can use Cisco Meeting App to run their meetings and perform actions such as viewing participants, muting and/or removing participants, starting and stopping recording, as well as creating and editing their own Spaces.

Note WebRTC App runs on only certain compatible browsers. For details on supported browsers, refer to the information at https://kb.acano.com/content/37/4/en/what-versions-of-browsers-do-we-support-for-webrtc-app.html.

Note Cisco Meeting App can be deployed inside or outside of the enterprise network to join a conference. For more deployment details, refer to the Cisco Meeting Server configuration guides available at https://www.cisco.com/c/en/us/support/conferencing/meeting-server/products-installation-and-configuration-guides-list.html.

Using Cisco Meeting Server for conferences has several benefits, including:

- Scaling easily for small or large deployments, allowing capacity to be added incrementally as needed
- Simplified, intuitive, and optimal conference experiences across all device types
- Unrestricted number of participants in a meeting, up to the limit of available underlying hardware when using multiparty licensing
- Single deployment model for all conference types

Role of Cisco TelePresence Management Suite (TMS)

Cisco TelePresence Management Suite (TMS) provides conference scheduling as well as conference room system reservations. Cisco Unified CM maintains the configuration control for endpoints, and Cisco TMS is then able to push the calendar information to those endpoints. Administrators are able to set the parameters for the default conference for their organization, and then individual conferences can be created according to this template.

Role of Cisco TelePresence Management Suite Extension for Microsoft Exchange (TMSXE)

When end users schedule a meeting in Microsoft Outlook with multiple conference room resources, the Exchange Web Services (EWS) feature of Microsoft Exchange synchronizes that event with Cisco TMS as a scheduled conference. This synchronization is bidirectional, allowing an administrator or support staff to update meetings as well without the need to access the meeting organizer's Outlook event. All endpoint resources within the organization that are intended to be in the conference must be listed in a single Microsoft Exchange meeting request.

Role of Cisco Meeting Management

Cisco Meeting Management is a standalone tool that does not require Cisco TMS, but together with Cisco TMS they provide the complete management functions for Cisco Meeting Sever. Cisco Meeting Management has the Meeting Manager that can provide white glove service for customers. The Meeting Manager lists all active meetings, with full details on each meeting. Within a particular meeting, it lists all participants in the meeting and can start/stop recording or streaming, change layout, add/drop participants, and end the meeting. For an individual participant, it can mute/unmute audio/video, change layout, and display call statistics.

Cisco Meeting Server Edge

Cisco Meeting Server Edge allows external users to join conferences from the Internet utilizing Cisco Meeting App. In this deployment, the external users interact with the edge software components that in turn communicate with the core software components. Edge components reside in the DMZ while core components reside within the enterprise network. When an external user connects, the call is routed to the edge components and then the core components. There are two options for deploying the edge components: single combined deployment (Figure 11-6) and single split deployment (Figure 11-7).

In a single combined deployment, the edge and core components are deployed inside the same server. However, the edge components are put into the DMZ network while the core components are put into the enterprise network inside the firewall. In a single split deployment, the edge and core components

are split into two different servers. The edge components are deployed on the server residing physically in the DMZ while the core components are deployed on a separate server residing inside the firewall in the enterprise.

Figure 11-6 Single Combined Deployment

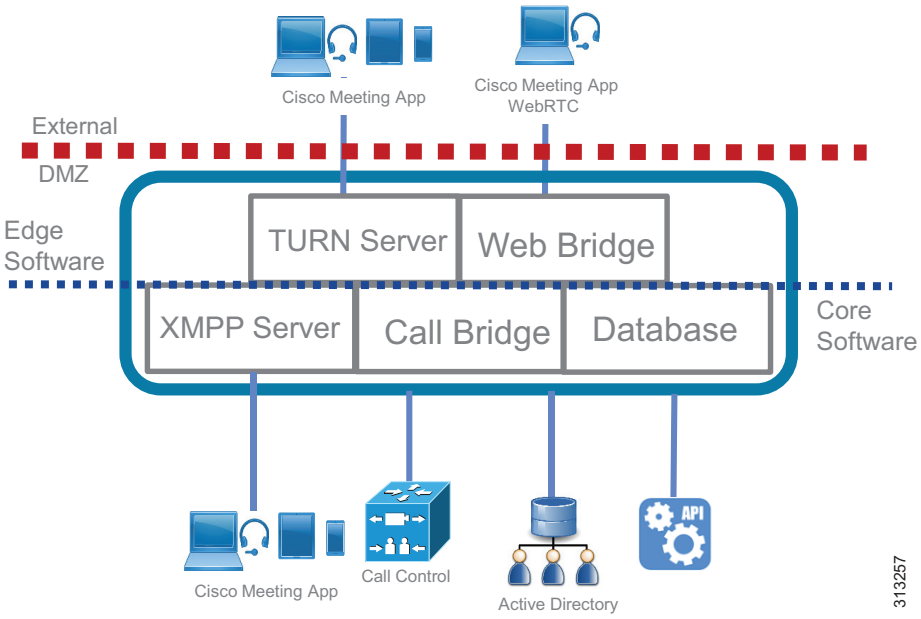
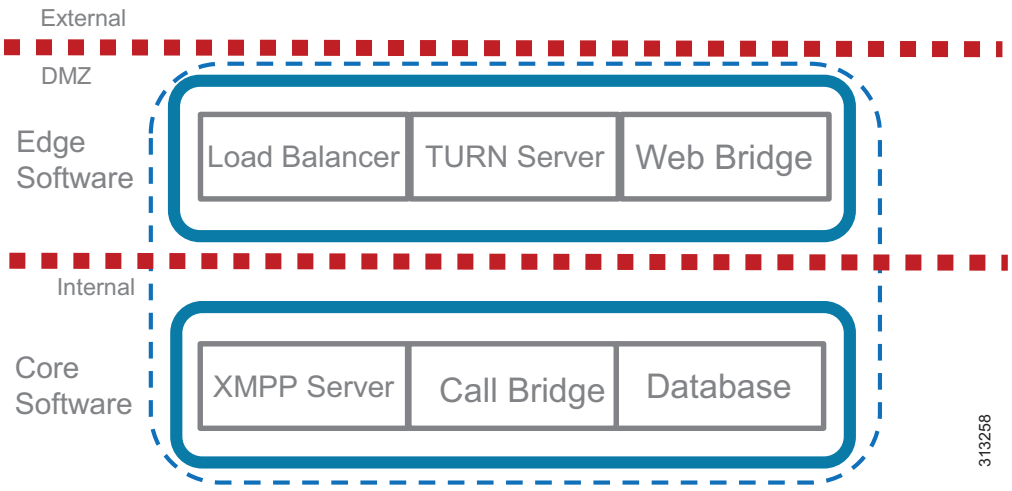


Figure 11-7 Single Split Deployment



External users using WebRTC clients connect directly to the Web Bridge in the DMZ. The TURN Server provides firewall media traversal technology that allows Cisco Meeting Server to be deployed behind a firewall or NAT. TURN is included with Cisco Meeting Server deployments without an additional license.

The Load Balancer provides a single point of contact for external Cisco Meeting App in the split deployment. It listens on an external interface and port for incoming connections. Also, the Load Balancer accepts incoming TLS connections from XMPP server, over which it can multiplex TCP connections from external clients. Thus, this creates a TLS trunk between the core and the edge components. The Load Balancer also does not require an additional license but it does require an enabled Call Bridge.

For more details on single combined deployments and single split deployments, refer to the respective deployment guides available at,

https://www.cisco.com/c/en/us/support/conferencing/meeting-server/products-installation-and-configuration-guides-list.html

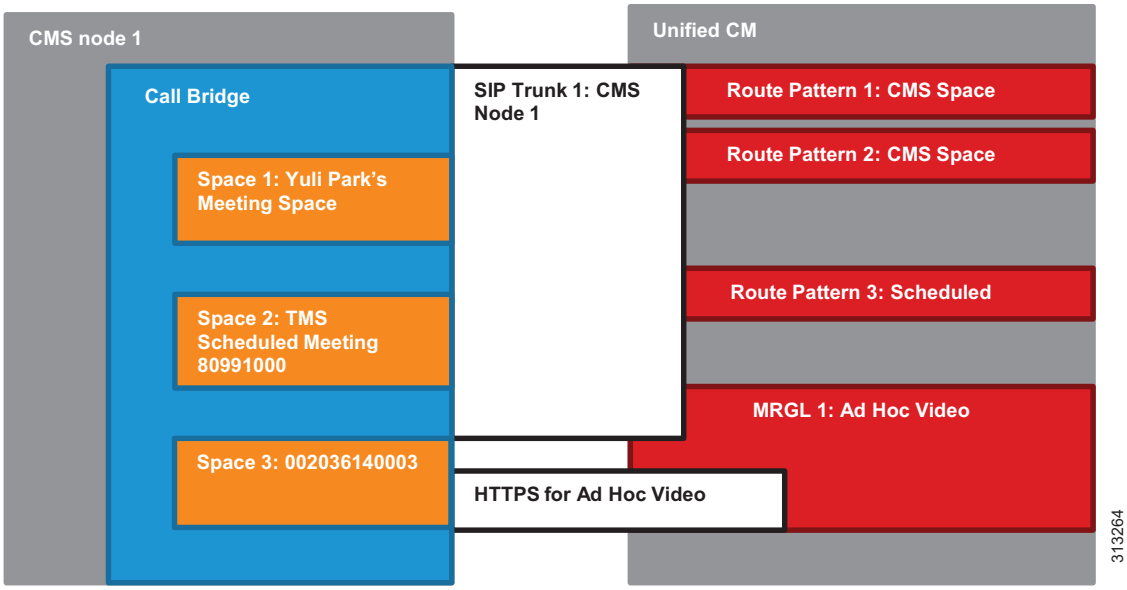
Starting with Cisco Meeting Server 2.1.2 and Cisco Expressway X8.10, external users using a WebRTC client can connect to Cisco Meeting Server over Expressway. For deployment details, refer to the latest version of the *Cisco Expressway Web Proxy for Cisco Meeting Server* deployment guide, available at

https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html

Conference Call Flows

Cisco Unified CM provides device registration and routing of voice and video calls between the connected endpoints. Permanent, instant, and scheduled conference calls are all routed over a single SIP trunk to the call bridge on Cisco Meeting Server. Each call bridge requires a separate SIP trunk. An HTTPS connection is configured on the Unified CM node that carries the XML-RPC requests to the Cisco Meeting Server nodes for instant conferences (see Figure 11-8). When users press the conference softkey on their device to escalate a two-party call to a three-party call, Unified CM sends an API request to Cisco Meeting Server to create a temporary Space for hosting the conference via this HTTPS connection. Instant, permanent, and scheduled conferences are hosted on Spaces that are created by various components.

Figure 11-8 Cisco Unified CM and Cisco Meeting Server SIP Trunk

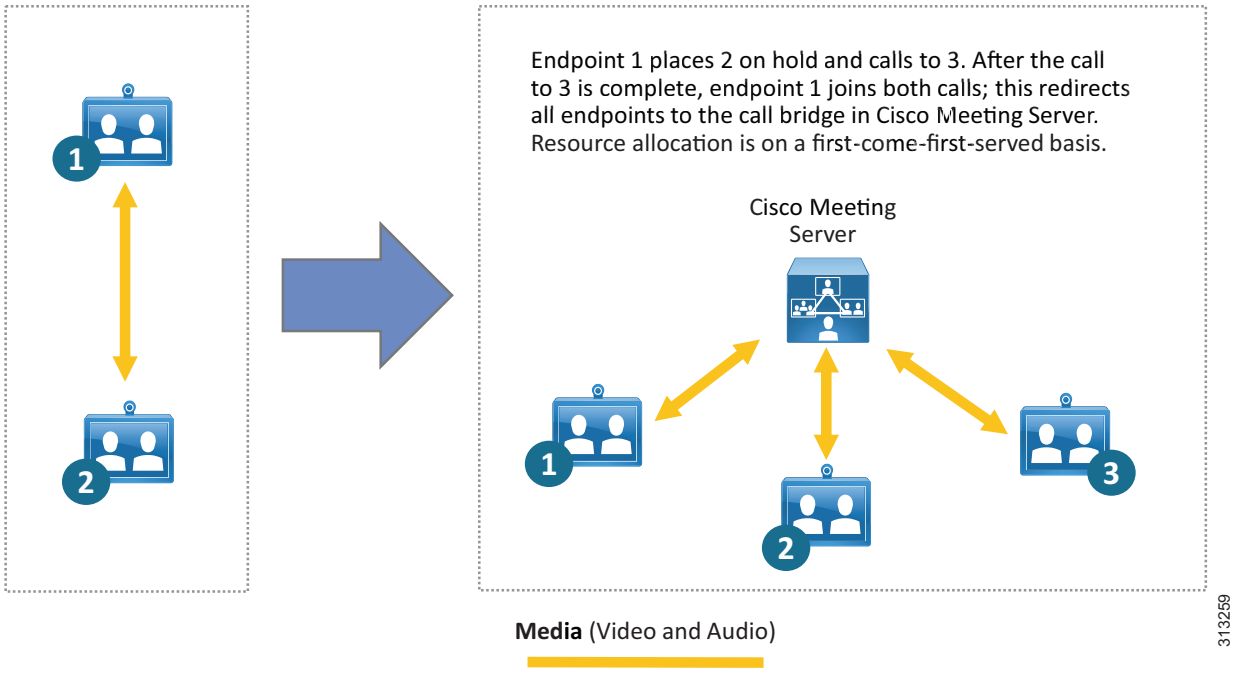


Instant call flows that are managed by Unified CM cannot be used to add participants to conferences created by any other method, such as scheduled conferences. Also, other call flows cannot be used to add participants to instant conferences. The instant call escalation method is supported only in an instant conference that was created by it, and conferences generated by other methods cannot be extended by the instant mechanism. This avoids any potential for chained conferences.

Instant Conferences

Instant conferences allow endpoints engaged in a two-party point-to-point call to be escalated to a three-party multipoint call using Unified CM conference bridge resources. Instant conferences use an HTTPS XML-RPC connection associated with the SIP trunk between Unified CM and the call bridge on Cisco Meeting Server. When a user presses the conference softkey to initiate an instant conference, Unified CM issues an API request through the HTTPS connection to create a temporary Space on Cisco Meeting Server. Unified CM then routes all the participants to that Space through the SIP trunk. When the conference ends, Unified CM issues another API request to delete that Space from Cisco Meeting Server. Figure 11-9 illustrates an example of how an instant conference is initiated using Cisco Meeting Server.

Figure 11-9 Initiation of an Instant Conference using Cisco Meeting Server



Permanent Conferences with Cisco Meeting Server Spaces

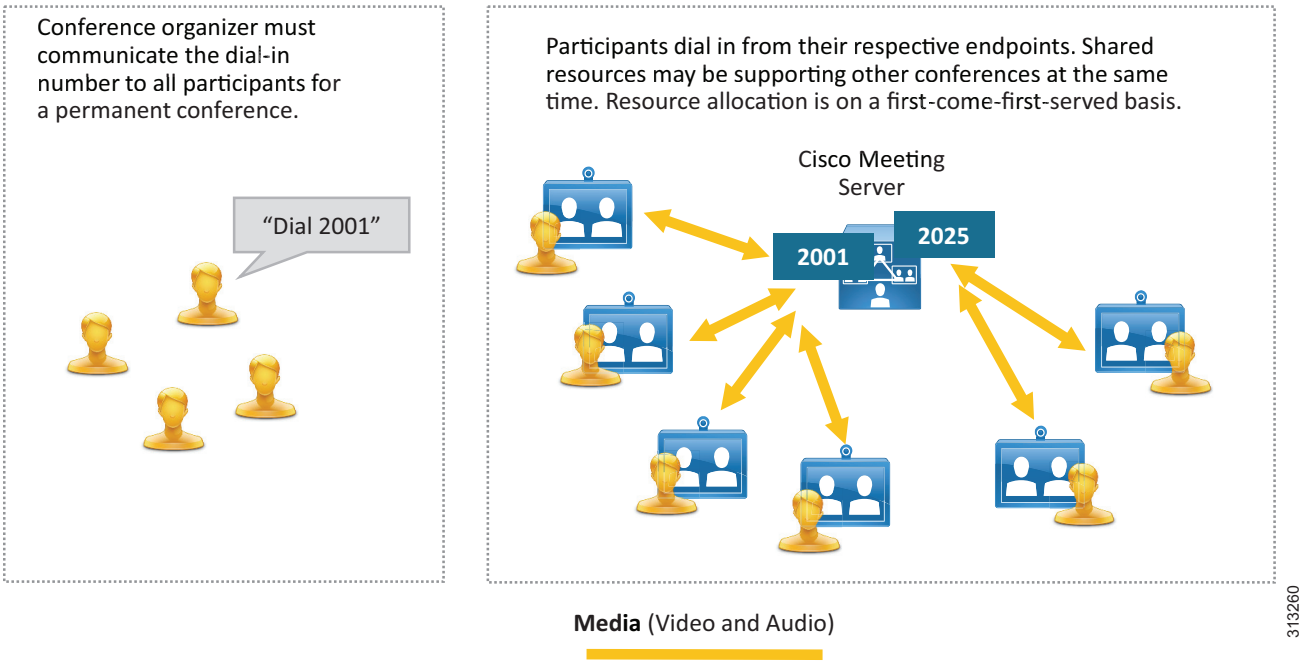
Permanent conferences are deployed using Cisco Meeting Server Spaces. Spaces provide a permanent-type conference and are created as part of the process of importing users from LDAP. Each Space has an associated video address (URI and/or DN) that users utilize to join the meeting. Administrators can specify the Space's attributes (for example, name, username, URIs, and so forth) through the field mappings so that the Spaces can be created using those mappings. Users can then log

in using Cisco Meeting App and add members to their Space. Also, users can log into Cisco Meeting App to create additional Spaces and add members. This conference type requires a SIP trunk between Unified CM and the call bridge on Cisco Meeting Server.

Permanent conferences can be initiated in several ways depending on the call control used by the conference initiator: IVR dial-in and preconfigured video address. Figure 11-10 illustrates an example of a permanent conference taking place by dialing a video address.

The permanent video address for a conference is always available after it has been preconfigured by the administrator, and the user can dial the video address to start or join the conference.

Figure 11-10 Example of a Permanent Conference



Alternatively, users can dial in to the conferences via the Cisco Meeting Server built-in interactive voice response (IVR). Then, the IVR will prompt the users for the conference ID and the password (if one is configured) of the conference they want to join.

Scheduled Conferences

Cisco Meeting Server supports scheduling conferences with Cisco TMS. Scheduled conferences require a SIP trunk between Unified CM and the call bridge on Cisco Meeting Server. Unified CM routes the scheduled conference participants to the destination of the SIP trunk. Add Cisco Meeting Server to Cisco TMS to allow for issuing REST API requests on Cisco Meeting Server through the HTTPS connection. After the administrator configures a range of numeric IDs for scheduled conferences, Cisco TMS creates an inactive Space on Cisco Meeting Server for each numeric ID via the API link. Cisco TMS then randomly chooses a dial-in number from the range when an organizer schedules a meeting. When it is time to start the scheduled meeting, Cisco TMS activates the Space using the API so that the participants can join. Scheduled conferences can be joined in a variety of ways, as Table 11-3 describes.

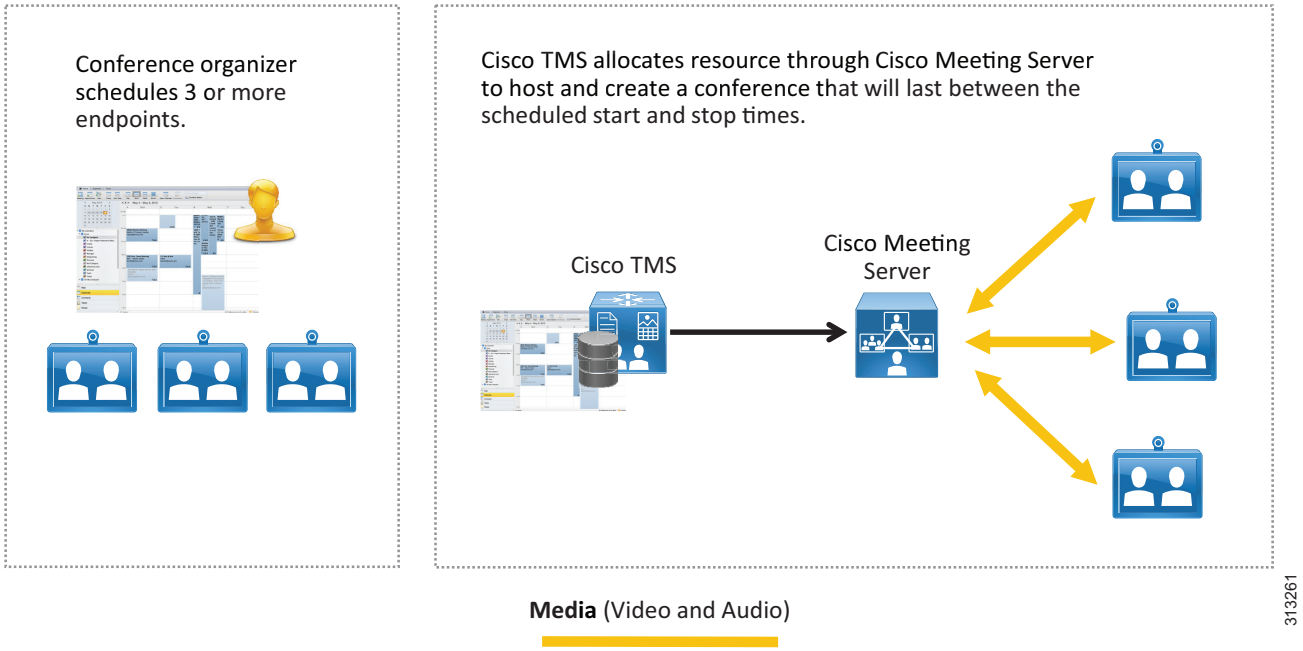
Table 11-3 Call Launch Options for Scheduled Conferences

Launch Method	Description
One Button to Push (OBTP)	• Conference dial-in information is automatically displayed on endpoints that support OBTP. For systems that do not support OBTP, an email with conference information is sent to the conference owner to forward to the participants.
Automatic Connect	• All endpoints are automatically connected at the specified date and time.
Manual Connect	• The conference cannot begin until a specific endpoint (usually the conference organizer's endpoint) connects. After this endpoint connects, the remaining endpoints are either connected automatically or allowed to dial in manually.
Reservation	• This method reserves the endpoints but does not initiate any connections.

Scheduling attempts to ensure endpoint and port resource availability and provides convenient methods to connect to the conferences. Most organizations already use calendaring applications to schedule conferences. In this case, calendar integration enables users to schedule conferences with their existing calendar client. TelePresence deployments often include a large quantity of endpoints and various infrastructure components. Without centralized management, component provisioning and resource allocation are difficult if not impossible. Management platforms greatly simplify these processes.

Scheduled meetings work by integrating conference resources and endpoints with corporate calendaring applications (see Figure 11-11). Cisco TelePresence Management Suite (TMS) resides between endpoints and calendaring applications to locate the proper bridge resource for each scheduled conference. Cisco recommends deploying scheduled conferencing with the TelePresence Management Suite and creating conferences by scheduling three or more endpoints.

Figure 11-11 Scheduled Conference Using Integration with a Calendaring Application



Security for Conferencing

Cisco Unified CM supports secure conferencing with Cisco Meeting Server by using secure SIP trunks between them. With secure conferencing, Unified CM uses TLS for call signaling and SRTP for media payload encryption. However, the entire conference is secure only if all the participants' endpoints support video encryption. The API interface between Cisco Unified CM and Cisco Meeting Server must be encrypted, and therefore HTTPS must be used for this.

Note Cisco Meeting Server 2.3 and later releases use TLS version 1.2 for all connections by default.

Cisco Meeting Server uses secure connections to communicate with external components as well as between internal components, and certificates are required. Both self-signed and certificate authority (CA) signed certificates are supported, but CA signed certificates are recommended in the deployment. For details on certificates deployment and requirements, refer to the latest version of *Cisco Meeting Server Certificate Guidelines* document, available at

https://www.cisco.com/c/en/us/support/conferencing/meeting-server/products-installation-and-configuration-guides-list.html

Another level of security can be added to restrict access to the conferences themselves with PINs or passwords. Any scheduled conference or permanent conference can have a PIN set so that all participants are challenged to enter the PIN before being allowed to connect.

For more information about secure conferencing, see the chapter on Cisco Collaboration Security, page 4-1.

High Availability for Conferencing

High availability must be considered at several levels with the conferencing solution and is achieved in different ways depending on the service being considered.

For both scheduled and non-scheduled conferences, high availability involves Cisco Unified CM, Cisco Meeting Server, and Cisco TMS.

Cisco Unified CM High Availability

To provide conference services with high availability, the link between Cisco Meeting Server and Cisco Unified CM must be redundant so that, if the link to one Cisco Meeting Server node goes down, the backup link can provide the services. These links include media resource groups and lists for instant conferences, and route groups and route lists for routing calls to Cisco Meeting Server.

Media Resource Groups and Lists

When a user of a Cisco Collaboration endpoint that uses Cisco Unified CM as its call control activates the CONF softkey, Unified CM uses the Media Resource Manager to select conference bridges. Conference bridge resources are configured in the media resource groups (MRGs). The media resource group lists (MRGLs) specify a prioritized list of MRGs and can be associated with the endpoints. The Media Resource Manager uses MRGLs of the endpoints to select the conference bridge. How you group

the resources is completely at your discretion, but Cisco recommends grouping the resources by using a logical model of the geographical placement whenever possible so that all endpoints at a given site use the conference bridges closest to them.

Cisco Unified CM selects conference bridges based on the following criteria, in the order listed here:

1. The priority order in which the media resource groups (MRGs) are listed in the media resource group list (MRGL)
2. Within the selected MRG, the resource that has been used the least

For further information about media resource design, see the chapter on Media Resources, page 7-1.

Route Groups and Lists

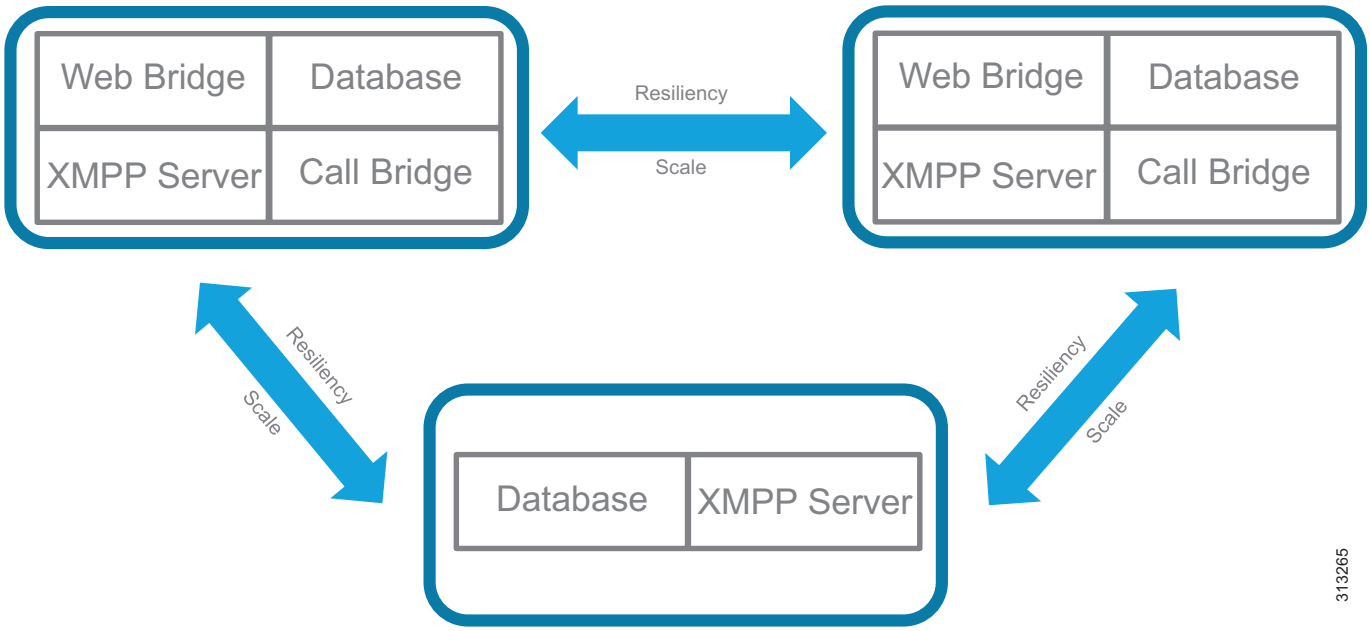
Route lists and route groups are common call routing mechanisms of reliability for calls that leave the Cisco Unified CM domain. For media resources integrated with Cisco Unified CM as a trunk, route lists and route groups should be used to achieve high availability if backup conference bridges exist. Call admission control can be preserved by setting the locations of the media resources based on the trunk being used for the call.

To learn more about route list and route group resiliency mechanisms, see the chapter on Dial Plan, page 14-1.

Cisco Meeting Server High Availability

Deploying additional instances of components on one or more servers can provide resiliency for Cisco Meeting Server so that the component instances can share the load, and if one of them fails, the backup instance would pick up the load. In addition, XMPP servers, call bridges, and databases can be clustered together to operate as a single instance, as shown in Figure 11-12.

Figure 11-12 Minimum Configuration for Cisco Meeting Server Cluster with High Availability



A standard Cisco Meeting Server cluster consists of two or more (up to 8) nodes with call bridge service enabled. Maximum round trip time (RTT) between nodes is 300 ms. Call bridge cluster peers are connected to each other in full mesh via the distribution link. This link is an HTTPS connection used for passing call signaling and control status messages between nodes. Calls can be sent to any nodes in the cluster. If one node goes down, Unified CM can route calls to the remaining call bridge nodes to join the conferences. In the event that a call bridge fails during a live conference, those calls will be dropped and participants will need to dial the same number to join the conference on a new call bridge. Using the Unified CM route group and route list construct, calls can be distributed through the SIP trunks to Cisco Meeting Server.

The database cluster consists of one master and multiple slaves, up to a maximum of 5 nodes with maximum RTT of 200 ms between nodes. The database master can perform both read and write operations, while slaves can only read. Call bridges always connect to the database master for read and write operations, and all changes made on the master are replicated to the slaves. Call bridges with a local database automatically connect to the master of the local database cluster, while call bridges with no local database have to be connected manually to the database cluster. If the master fails, one of the slaves will become the new master, and other slaves will re-register with this new master. After correcting the failure, the old master will become the slave and register with the new master. In cases where a network partition occurs, only database nodes that can see more than half of the cluster members are considered for promotion to become a master. Likewise, any existing master that cannot see more than half of the cluster members will be demoted to a slave. This ensures that multiple masters are not created. Thus, if a database cluster consists of an even number (2 or 4) of nodes and the network is partitioned into 2 segments with an equal number of nodes on each segment, the master on one side will be demoted to a slave because it cannot see more than half of the cluster members. In that case, there will be no master in the cluster, and the call bridges can still take calls but no database write operations will be possible. For this reason, we recommend having an odd number of nodes in the database cluster to ensure that a master is always elected. As a result, the minimum number of database nodes in a cluster is 3.

XMPP resiliency provides failover protection for a client that is unable to reach a specific XMPP server. The XMPP server cluster must be configured using an odd number of XMPP server nodes, with a minimum of 3 nodes. This is due to the master election algorithm requirement that more than half of the cluster nodes should be available in order for Cisco Meeting Server to elect an XMPP server master. If no XMPP server master is available in the cluster, Cisco Meeting App users cannot log in. Each XMPP server knows the location of the others, with links established between them. They use keep-alive messages to monitor each other and elect a master. XMPP messages can be sent to any server and are forwarded to the master XMPP server. If the master fails, a new master is elected and the other XMPP servers will forward messages to the new master. The call bridge uses the DNS SRV record (_xmpp-component) to connect with an available XMPP server based on the configured priority and weight with the SRV record. A call bridge connects to one XMPP server at a time. If a network problem results in the call bridge losing connection to the XMPP server, the call bridge will attempt to reconnect to another XMPP server. All call bridges must be configured inside each XMPP server.

Figure 11-12 illustrates the minimum configuration for a Cisco Meeting Server cluster with high availability. In this configuration, a minimum of 3 servers is required to host 3 instances of the database and XMPP servers. Enable at least 2 instances of each component service (Web Bridge and Call Bridge) in separate servers, and put the call bridges into a group. There is no need to activate all services inside each server; activate only the ones that are required. If the deployment requires more capacity than the two call bridges can handle, additional call bridge can be set up in the third server (no need to acquire a fourth server for just the call bridge).

Cisco TMS High Availability

High availability of a large Cisco TMS deployment includes two TMS front-end servers, two servers running TMSXE, a network load balancer, and an external Microsoft SQL database (see Figure 11-3). TMS resiliency supports only two servers (one active node and one passive node), and this model does not increase or decrease the capacity of the TMS deployment. The network load balancer (NLB) is deployed in front of the TMS servers. Inbound traffic to TMS goes through the NLB, which forwards it to the active node. Outbound traffic from TMS is sent directly to the destination without going through the NLB. If the NLB detects a failure on the existing active node, it automatically switches to the new active node without any user intervention.

Cisco Meeting Management High Availability

Cisco Meeting Management does not have the built-in cluster functionality to provide resiliency. However, if customers want to have high availability, they can configure two individual Cisco Meeting Management instances with exactly the same configuration managing the same Cisco Meeting Server instance(s), and they can then put a network load balancer in front of the two Cisco Meeting Management instances. Users can then connect to the Cisco Meeting Management portal through the load balancer. The load balancer configuration and availability of the Cisco Meeting Management server will determine which one of the Cisco Meeting Management servers will be used.

Scaling the Conferencing Solution

The conferencing solution can be scaled by adding more call bridges (up to 8) to a standard Cisco Meeting Server cluster.

In this deployment, based on the dial plan and route group and route list configuration with the SIP trunks in Unified CM, calls can be routed to any call bridge within the cluster. If calls for the same conference are routed to different call bridges, the audio and video of the last 4 active speakers are exchanged between call bridges so that participants on one bridge can see the active speakers on the other bridge.

Note Cisco Meeting Server supports clustering with more than 8 call bridges, but deployment requires prior approval by Cisco. Contact your local Cisco account team for details.

Each call bridge can support 450 participants. Thus, the maximum number of participants per conference is 450 with a single server, and up to 2,600 participants can be supported across multiple servers in a single cluster. Additional scalability information can be found in the latest version of *Cisco Meeting Server and Cisco Meeting App Data Sheet*, available at

https://www.cisco.com/c/en/us/products/conferencing/meeting-server/datasheet-listing.html

Using call bridge groups, a Cisco Meeting Server cluster can intelligently load balance calls across the call bridges within the same location or across nodes in different locations to increase the scale within a deployment. The intelligent decision making to determine where calls end up is handled by Cisco Meeting Servers. The call control system needs to be able to handle the SIP messages from Cisco Meeting Servers in order to move calls to the correct location. Call bridges that are configured as a cluster can be put into one or more call bridge groups. For call bridges within the group, Cisco Meeting Server intelligently load balances calls across them and sends calls for the same conference to the same call bridge whenever possible in order to minimize the creation of distribution links between call bridges.

For inbound SIP calls to a call bridge, Cisco Meeting Server decides to reject or accept the call based on the current load in the call bridge. If the current load is less than the preset threshold, the call will be accepted. Otherwise, the call will be rejected and Unified CM will reroute the call to another call bridge in the call bridge group using the dial plan configuration. If Unified CM cannot find any call bridge that accepts the call, the whole call will be rejected. After a Cisco Meeting Server accepts the call, the call could be hosted on the call bridge of this Cisco Meeting Server or moved to another call bridge with highest priority according to an internal ordered list for the conference. When the call is moved, the target Cisco Meeting Server with the call bridge enabled sends an INVITE with Replaces to Unified CM to take over the call. By default, a call bridge in a call bridge group will reject all calls for new participants at 80% load, and only new distribution calls will be allowed.

For outbound SIP calls, Cisco Meeting Server uses the outbound dial rule to locate the highest priority rule that matches the domain, and it load balances the call using a call bridge group. If the rule applies to a local call bridge, calls will be load balanced using the local call bridge group. Otherwise, calls will be load balanced using the call bridge group where the remote call bridge is a member. Also, for an outbound call made using an API, a call bridge group or a call bridge can be specified as a parameter. For a call bridge group, the call is load balanced among the call bridges within that group. For a call bridge, the call is made using that call bridge.

For Cisco Meeting App (including WebRTC) clients, users can join the conference as members of the Space, as non-members of the Space with accounts on Cisco Meeting Server, or as guests. When a user is added as a member to the Space via an API, a call bridge group or a call bridge can be specified as a parameter. For a call bridge group, when the Cisco Meeting App user joins the meeting, the call will be load balanced using that group. For a call bridge, when the Cisco Meeting App user joins the meeting, the call will be handled by that call bridge. When the user is not a Space member or is a guest that joins the meeting via Cisco Meeting App, the first call bridge that the user connects to is determined. If that call bridge is part of a call bridge group, then the call is load balanced using that group. In order to load balance Cisco Meeting App calls, ensure that each call bridge in the call bridge group has a connection to the XMPP cluster or to a single XMPP server.

There are some network requirements that must be satisfied in order to successfully move calls between call bridges within the call bridge group. The maximum RTT should be 100 ms between call bridges inside the group and 300 ms between any two call bridges within the cluster.

For setup and deployment details for call bridge groups, refer to the latest version of the white paper on *Load Balancing Calls Across Cisco Meeting Servers*, available at,

https://www.cisco.com/c/en/us/support/conferencing/meeting-server/products-installation-and-configuration-guides-list.html

Note If call bridge groups and load balancing are not used, then calls will not be rejected but the quality of all calls will be reduced when the load limit is reached. If this happens often, we recommend deploying additional hardware.

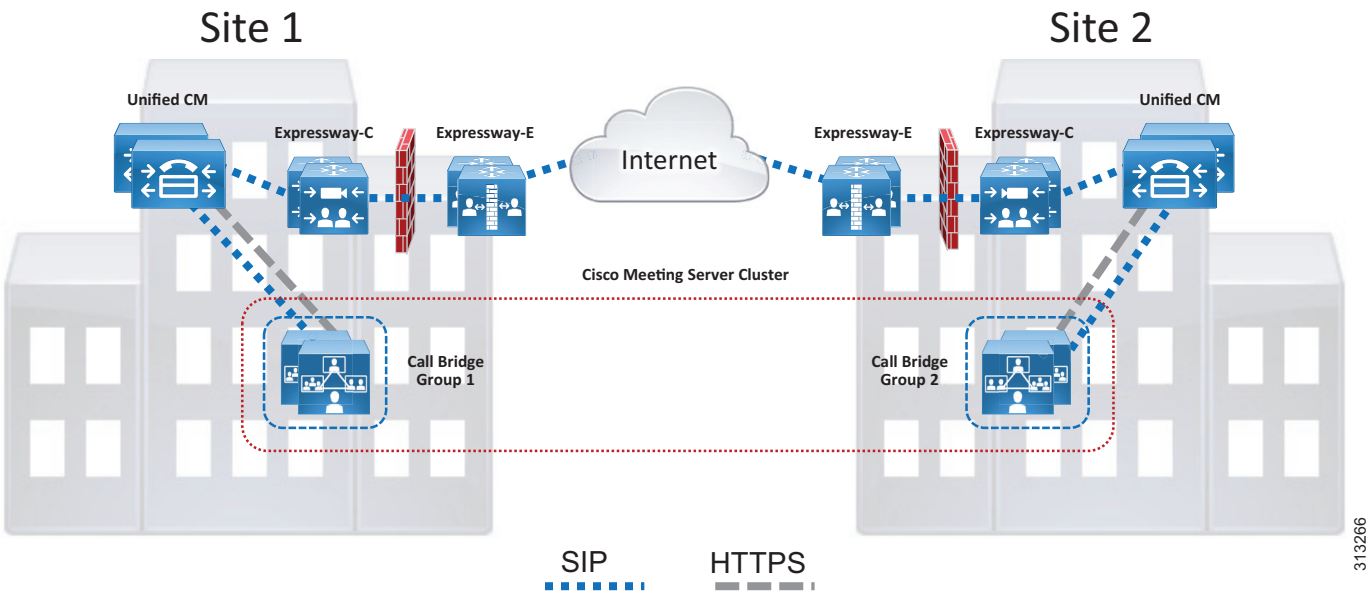
Considerations for Multiple Unified CM Clusters

For large deployments with multiple Cisco Unified CM clusters, use a single Cisco Meeting Server cluster configured with multiple call bridge groups, and dedicate one group to each Unified CM cluster.

For example, if your deployment has three Unified CM clusters, then you should deploy a single Cisco Meeting Server cluster with three call bridge groups, one for each Unified CM cluster. Each Unified CM cluster should have a SIP trunk to each call bridge in its local call bridge group. All incoming conference calls to a Unified CM cluster will be handled by the local call bridge group. Call bridges should have their distribution links connected to their peers inside and outside of the groups in full mesh. For the

same conference, users can dial in from their Unified CM cluster to reach the local call bridge group, and the call bridges in different groups will exchange the audio and video of the last 4 active speakers with their peers so that participants can see each other across the bridges. (See Figure 11-13.)

Figure 11-13 Cisco Meeting Server Deployment with Multiple Unified CM Clusters



The following guidelines apply when expanding the Cisco Meeting Server cluster into different regions for multiple Unified CM clusters:

- A single Cisco Meeting Server cluster should be used for deployment of one or more Cisco Unified CM clusters.
- You may deploy up to 8 call bridges for the standard Cisco Meeting Server cluster. If the cluster exceeds 8 call bridges, acquire Cisco account team approval before deployment.
- Deploy a maximum of 5 databases and an odd number of nodes in the Cisco Meeting Server cluster.
- Deploy an odd number of XMPP service nodes in the Cisco Meeting Server cluster.
- Round-trip-time (RTT) network requirements:
  - Maximum of 300 ms between call bridges and 200 ms between databases in the Cisco Meeting Server cluster
  - Maximum of 100 ms between call bridges inside the group

Licensing

Cisco Meeting Server supports both Multiparty licensing and Cisco Meeting Server Capacity Units, but Multiparty licensing is recommended. Multiparty is a user-based licensing model, and it should be applied to every node that has the call bridge enabled. It comes with two variations: Personal and Shared. Personal Multiparty Plus (PMP+) is for specific named hosts, while Shared Multiparty Plus (SMP+) is for conference room systems or for sharing between users. By default, all users in the system use Shared Multiparty Plus (SMP+); and if Personal Multiparty Plus (PMP+) is desired, PMP+ should be assigned to users via the Cisco Meeting Server API. Each license entitles a user to host a conference with unlimited participants and up to 1080p video resolution. Table 11-4 summarizes the features included in the Personal and Shared Multiparty licenses.

Table 11-4 Cisco Personal and Shared Multiparty Plus License Features

Feature	Personal Multiparty Plus (PMP+)	Shared Multiparty Plus (SMP+)
Tied to a named host	Yes	No
Availability	Included in Cisco UWL Meeting	A la carte or discounted with room system
Minimum order	25	1
Maximum conference size	Unrestricted, within the limit of available hardware capacity	
Maximum resolution	1080p60 (full HD) for video and 1080p30 for content on single-screen or multi-screen endpoints	
Rich media sessions for business-to-business or business-to-customer	Included	Included
Cisco TMS, TMSXE, and Skype for Business and Lync Interoperability Licenses	Included	New customers buy with Starter Pack[1]
Support for instant, permanent, and scheduled conferences	Yes	Yes

1. If only the Cisco TMS and related product licenses are required, a TMS Starter Pack can be purchased.

For more information on licensing, refer to the Cisco Meeting Server product documentation available at

https://www.cisco.com/c/en/us/products/conferencing/meeting-server/index.html?dtid=osscdc000283

Cisco Meeting Management does not require any additional license for Cisco Meeting Server customers, and it can be downloaded from https://www.cisco.com.

Capacity Planning

The capacity of Cisco Meeting Server depends on the platform of choice and the number of conferencing nodes running in the deployment. The main purpose of sizing a deployment is to determine the number of required concurrent connections to the Cisco Meeting Servers. Considerations include:

- Geographical location — Each region served by Cisco Unified CM should have dedicated conferencing resources. For example, there could be one central location for the US where Unified CM, Cisco Meeting Servers, and other servers are installed, and one central location for EMEA.
- Preference for Cisco Meeting Server platforms — Virtualized or appliance
- Cisco Meeting Server platform capacities — For capacity details, refer to the Cisco Meeting Server data sheet available at https://www.cisco.com/c/en/us/products/conferencing/meeting-server/datasheet-listing.html.
- Type of conferences — Audio and/or video; scheduled and/or non-scheduled
- Conference video resolutions — Higher quality conferences use more resources.
- Large conference requirements — For example, all-hands meetings

Conference resources are generally dedicated to a region in order to keep as much of the conference media as possible on the regional network; therefore, sizing can be considered on a region-by-region basis. To properly size the conferencing deployment, use the Cisco Collaboration Sizing Tool, available (with a valid login account) at https://cucst.cloudapps.cisco.com.

Note Consult with your Cisco sales representative for assistance on sizing the conferencing resources for your particular environment.

For additional sizing details, see the section on Collaborative Conferencing, page 25-44.

Depending on the Cisco Meeting Server platform types and the number of Cisco Meeting Server clusters to be managed by Cisco Meeting Management, different deployment size will be required. For details, refer to the latest Cisco Meeting Management release notes, available at

https://www.cisco.com/c/en/us/support/conferencing/meeting-management/products-release-notes-list.html

Design Considerations

In summary, consider the following recommendations when deploying Cisco Meeting Server:

- Use Multiparty Licensing (PMP+ and/or SMP+) for Cisco Meeting Server deployments.
- Cisco Meeting Server deployments require certificates to secure both internal and external connections and should use the CA signed certificates.
- The Cisco Meeting Server cluster supports up to 8 nodes with call bridges. More than 8 nodes (up to 24) can be deployed, but this requires Cisco approval before deployment.
- When the deployment involves clustering the database or XMPP service, a minimum of 3 databases or XMPP nodes is required due to the master selection algorithm requirement for clustering.
- In a large distributed Cisco Meeting Server deployment, use call bridge groups to group the call bridges together within the region to minimize the creation of distributed links between call bridges.
- Network requirements for Cisco Meeting Server cluster deployment are 300 ms between call bridges, 200 ms between databases, and 100 ms between call bridges inside a call bridge group.
- Cisco TelePresence Management Suite (TMS) provides conference scheduling and endpoint management, while Cisco Meeting Management provides meeting management. These two applications provide the complete management solution for Cisco Meeting Server.
- Cisco Meeting Management can be deployed without an extra license, but it requires a separate server from Cisco Meeting Server.

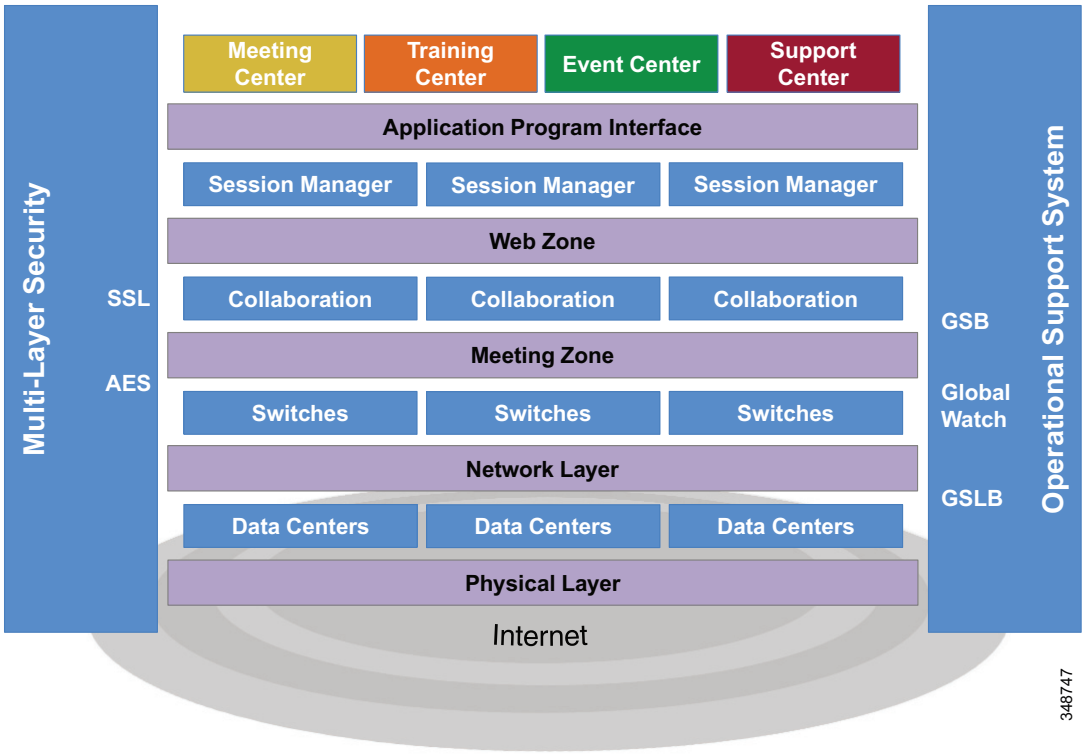
Cisco WebEx Software as a Service

Cisco WebEx is a collaborative conferencing solution that does not require any hardware to be deployed on-site. All services (audio, video, and content sharing) are hosted in the Internet through the Cisco WebEx Collaboration Cloud. This is often referred to as software-as-a-service (SaaS). Meetings can be initiated and attended from anywhere, anytime, on any device, and do not require connectivity back into the enterprise apart from an Internet connection. This section describes solution characteristics and provides design guidance for deploying WebEx SaaS.

Architecture

Cisco WebEx SaaS utilizes the Cisco WebEx Collaboration Cloud to deliver the conferencing solution to the customers. The Cisco WebEx Collaboration Cloud is a global network created with a carrier-class information switching architecture, and only Cisco Collaboration traffic flows over this network. Figure 11-14 shows the Cisco WebEx Collaboration Cloud architecture.

Figure 11-14 Cisco WebEx Collaboration Cloud Architecture



This network is purpose-built for real-time communications and has been specially formulated to minimize latency associated with TCP-layer flows. The network consists of application-specific multimedia switches at key peering points to handle rapid session traffic and to guarantee a high quality of service for WebEx meetings. These switches are housed in highly secure Cisco data centers interconnected via dedicated lines that circumvent the public internet. These data centers are located near the major internet access points to route meeting traffic around the globe securely and reliably. In addition to these large data centers housing major meeting nodes, Cisco deploys nodes around the world. The network is built on fully redundant clusters with Global Site Backup. These services and other facilities form part of the Cisco WebEx Collaboration Cloud Operational Support System.

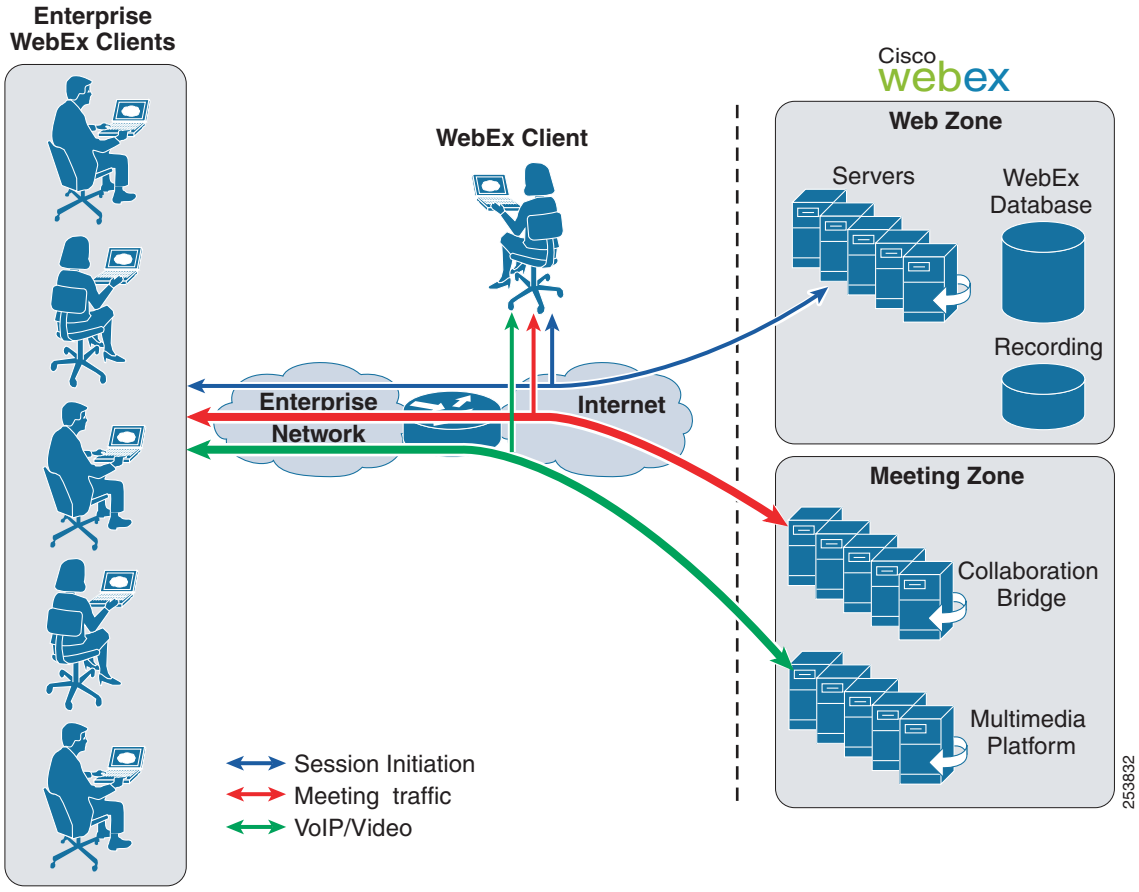
Users can connect to a WebEx meeting using the meeting application running on a computer or mobile device or even an HTML5 browser-based web application. Once the connection is established, the WebEx Collaboration Cloud manages all synchronous real-time interactions that make up a WebEx meeting, as depicted in Figure 11-14. Users access WebEx applications via browsers through the WebEx Collaboration Cloud, which resides within the Web Zone. The Applications Program Interface (API) ties the WebEx applications to the switching platform in the Meeting Zone within the WebEx Collaboration Cloud core. Numerous clusters of interconnected and distributed collaboration switches, their associated databases, and the logical and physical network infrastructure make up the WebEx Collaboration Cloud core. Multi-layer security components and the WebEx Operational Support System encircle the network with an additional layer of protection.

The WebEx Collaboration Cloud delivers real-time traffic reliably using intelligent routing, Global Site Backup (GSB), and Global Server Load Balancing (GSLB). Based on the geographic location of WebEx meeting participants, the WebEx Collaboration Cloud determines the point of presence that offers the lowest latency and best performance. WebEx meeting hosts automatically get a backup site physically located in a geographically distant Cisco data center within the same region. In the unlikely event that the primary WebEx site becomes unavailable, GSB automatically switches all meeting activity to the

backup site. GSLB is a load-balancing design that directs traffic to the least congested switch in the WebEx Collaboration Cloud in order to minimize the delays. Thus, if one meeting switch has congestion, traffic is directed to an alternate switch, resulting in faster screen updates and synchronization among participants, and a better meeting experience.

In the WebEx deployment model shown in Figure 11-15, all the content, voice, and video traffic from every client traverses the internet and is mixed and managed in the cloud at the WebEx data center. The WebEx data center is logically divided into the Meeting Zone and the Web Zone. The Web Zone is responsible for things that happen before and after a web meeting. It incorporates tasks such as scheduling, user management, billing, reporting, and streaming recordings. The Meeting Zone is responsible for switching the actual meeting once it is in progress between the endpoints.

Figure 11-15 WebEx Deployment



The Meeting Zone consists of two subsystems. Within the Meeting Zone there are collaboration bridges that switch meeting content. The multimedia platform is responsible for mixing all of the VoIP and video streams within a meeting. To join a WebEx session, an attendee first connects to the Web Zone. The Web Zone traffic flows only before or after the meeting, is relatively low bandwidth, and is mainly non-real time. The real-time meeting content share flows to and from the Meeting Zone and can be bandwidth intensive. Its real-time nature can place a heavy burden on enterprise access infrastructure. For further details regarding network traffic planning, see Capacity Planning, page 11-33.

Meeting Center uses the H.264 AVC/SVC codec to provide high-definition video for the conference. Higher network bandwidth is needed for those deployments. For further details regarding network traffic optimization for high-definition video, see Capacity Planning, page 11-33.

Each WebEx Meeting Center host has a Personal Room with a fixed customizable URL. The host can use his room to conduct meetings, and participants can enter the room using that fixed URL. Lobby management functions are available to the host, such as maintaining privacy by locking the room to prevent others from entering while the meeting is in progress.

Starting with Cisco WebEx Meeting Center version WBS32.4, Meeting Center supports audio noise detection functionality that can distinguish between users who are actively speaking and background noise. When audio is connected via the Call Using Computer option, Meeting Center can automatically warn users and prompt them to mute if active background noise (such as knocking, typing, a siren, or dogs barking) is detected. However, users have the option to disable the noise detection function if desired.

Note Audio noise detection supports the Call Using Computer audio option only.

Security

By default, all WebEx meeting data is encrypted using 128-bit SSL encryption between the client and Cisco's Collaboration Cloud. SSL accelerators within the cloud decrypt the content sharing information and send it to a WebEx conference bridge that processes the content and sends it back through an SSL accelerator, where it is re-encrypted and sent back to the attendees. All Web Zone and Meeting Zone traffic is encrypted using 128-bit SSL where SSL accelerators are used to off-load the SSL function from the Web and Meeting Zone servers.

After the meeting ends, no session data is retained in the WebEx cloud or an attendee's computer. Only two types of data are retained on a long-term basis: billing and reporting information and optionally network based recordings, both of which are accessible only to authorized enterprise users.

Some limited caching of meeting data is carried out within the Meeting Zone, and this is done to ensure that users with connectivity issues or who may be joining the meeting after the start time receive a current fully synchronized version of the meeting content.

Independent third parties are used to conduct external audits covering both commercial and governmental security requirements, to ensure the WebEx cloud maintains its adherence to documented security best practices. WebEx performs an annual SSAE 16 audit in accordance with standards established by the AICPA, conducted by Price Waterhouse Coopers. The controls audited against WebEx are based on ISO-27002 standards. This highly respected and recognized audit validates that WebEx services have been audited in-depth against control objectives and control activities (that often include controls over information technology and security related processes) with respect to handling and processing customer data.

For customers that require enhanced security, there is also an option to perform end-to-end 256 bit AES encryption for collaboration bridge and multimedia content so that traffic is never decrypted in the cloud. End-to-end encryption results in some lost features such as NBRs. For more information on enhanced WebEx security options, refer to the white paper *Unleash the Power of Highly Secure, Real-Time Collaboration*, available at

https://www.cisco.com/en/US/products/ps12584/prod_white_papers_list.html

Note End-to-end encryption options are available for Meeting Center and Support Center meetings without additional cost.

Site administrators can enforce the use of passwords to access the network-based recordings. When the host schedules the meeting, he/she can require all attendees joining the meeting to sign in with Single Sign-On (SSO) authentication and can restrict the meeting to invited attendees only. In addition, site administrators can enable an option to allow only authenticated attendees to enter a host's unlocked Personal Room, while the unauthenticated attendees must wait in the lobby until the host admits them to enter the room.

Scheduling

With respect to scheduling and initiating meetings, WebEx provides cloud-based web scheduling capability, but most organizations prefer to schedule from their corporate email system (Exchange, Lotus Notes, and so forth) or other enterprise applications. The WebEx Productivity Tools is a bundle of integrations with well known desktop tools incorporated into a single application. A WebEx administrator can control the specific integrations that are provided through the tool to their organization's user population. It can be downloaded and installed from the WebEx site, or it can be pushed out locally using standard desktop management tools. To learn more about WebEx Productivity Tools, refer to the information available at

https://www.webex.com/support/productivity-tools.html

User Profile

There are several options for creating WebEx user profiles for an organization in the cloud. Security considerations for the actual usernames and passwords, as well as for handling a large number of user accounts, should be considered. A WebEx administrator can create user profiles manually by bulk import of a CSV template or by a programmatic approach. A programmatic approach uses one or a combination of the WebEx APIs, URL, and XML, or a Federated SSO solution. The programmatic approach can be used by a customer portal, which is an application such as a CRM tool or a Learning Management System that integrates directly into WebEx. In addition, the user can sign up for an account from the company's WebEx site, and the user profile will be created after the request has been approved.

For integrating directly with an organization's LDAP directory, Federated SSO with Security Assertion Markup Language (SAML) is the preferred approach. For more information regarding Federated SSO, refer to the white papers and technical notes available at

https://developer.cisco.com/site/webex-developer/develop-test/sso/reference/

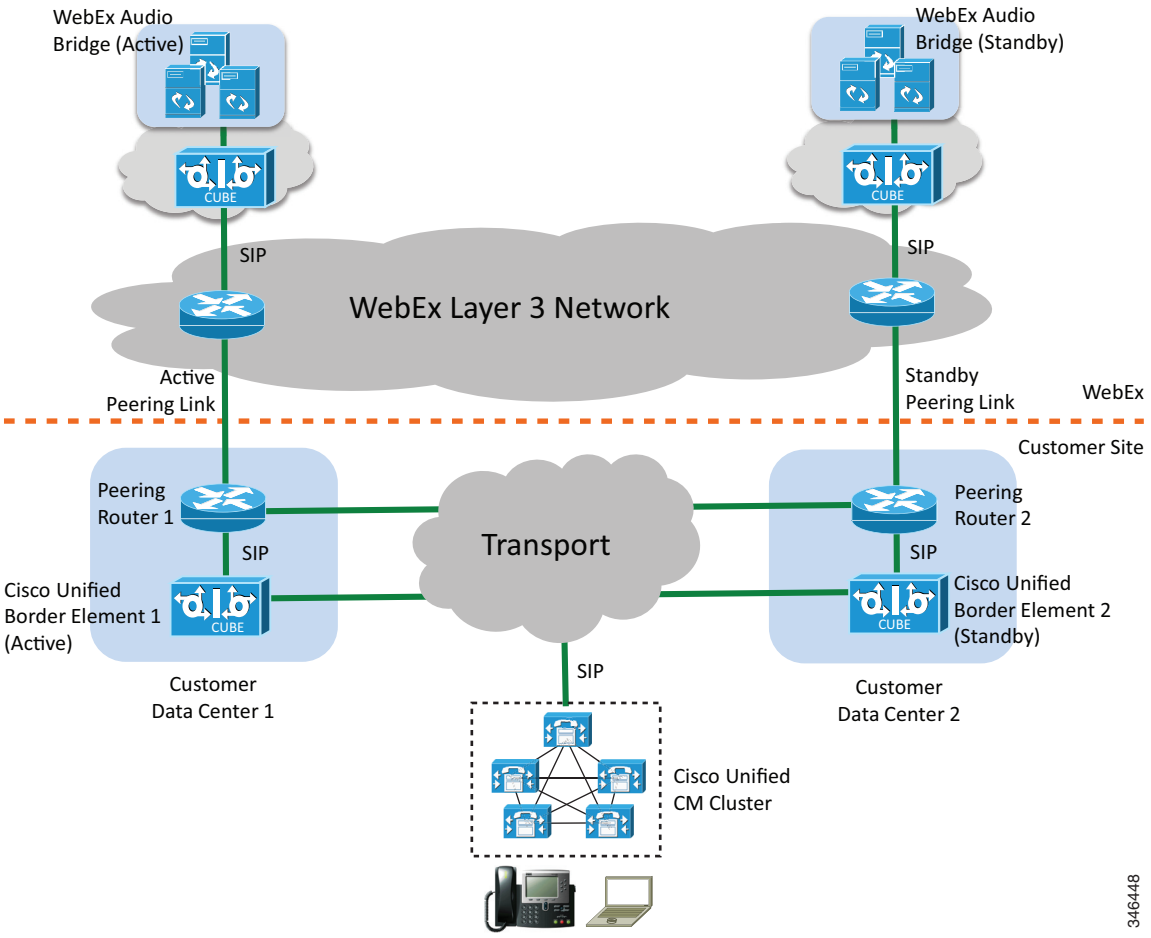
High Availability

The Cisco WebEx Collaboration Cloud has a very high level of redundancy built in and is managed by Cisco. It is designed for continuous service with a very robust cut-over to the redundant meeting nodes during outages. In addition to the primary WebEx site, every customer has a backup site physically located in a geographically distant WebEx data center within the same region. If a customer's primary site is unavailable, Global Site Backup (GSB) automatically moves all meeting activity to the backup site. Neither the hosts nor the participants notice that they are being redirected to the backup site. The GSB system facilitates continuous accessibility to WebEx meetings globally, and all attributes, address books, preferences, meeting schedules, and other real-time data are kept in sync between the primary and backup sites. Because of this synchronization, GSB provides redundancy and disaster recovery both before and after the meetings.

Cisco WebEx Cloud Connected Audio

Cisco WebEx Cloud Connected Audio (CCA) is an audio conferencing solution based on a hybrid deployment model that uses the on-premises IP telephony network to provide an integrated audio experience for an organization's WebEx meetings. WebEx CCA implements a SIP trunk connection from the organization's IP telephony network into the WebEx cloud infrastructure (see Figure 11-16). The audio conferencing traffic traverses through this SIP connection instead of the service provider PSTN connection and, thus, WebEx CCA provides significant savings on audio cost and maintains the same integrated and intuitive user experience as other WebEx audio options.

Figure 11-16 Cisco WebEx Cloud Connected Audio High-Level Design



As shown in Figure 11-16, a typical WebEx CCA high-level design consists of the on-premises IP telephony network and the WebEx cloud infrastructure that are connected via the dedicated IP Peering Connections provided by the customer. The on-premises IP telephony network consists of a Cisco Unified Communications Manager (Unified CM) cluster and Cisco Unified Border Element. Cisco Unified Border Elements are deployed in the WebEx cloud infrastructure and they mark the entry point for an organization's IP telephony network. The Cisco Unified Border Elements in the cloud and at the customer site communicate with each other via SIP. WebEx CCA requires the customer to have two IP Peering Connections that connect with different WebEx data centers residing in geographically separated locations for redundancy purpose. The redundant IP links are configured in active/standby

mode. All conferencing audio traffic flows through the primary link and fails-over to the secondary link if the primary link goes down. WebEx CCA also requires the gateway routers to support Border Gateway Protocol (BGP) and Bidirectional Forwarding Detection (BFD) protocol. BGP and BFD offer a significant faster re-convergence time in the event of a network failure.

Note The WebEx data center equipment, audio bridge, and servers run over the shared infrastructure along with other customers in the WebEx CCA solution.

Cisco Unified CM has a SIP connection with the WebEx cloud through the Cisco Unified Border Element at the customer site to handle telephony signal. The conference dial-in number is owned by the customer and is terminated at the customer site. Call routing is handled at customer the site, call signaling and audio traffic is handled over the redundant IP peering connections, and call mixing is handled in the cloud. When users dial the conference number within the enterprise, Cisco Unified CM routes the call over the dedicated SIP trunk through the Cisco Unified Border Element to the WebEx cloud without traversing through the PSTN. When the conference users request callback, WebEx sends the call to the Cisco Unified Border Element at the customer site that routes it to the destination end-point. If the conference users reside outside of the enterprise network, calls are routed through the PSTN before terminating or after leaving the customer's IP telephony network. WebEx CCA supports only the G.711 audio codec, RFC 2833 DTMF, and SIP signaling.

WebEx CCA has the highly available and fully redundant architecture that is designed to ensure continuous service operation. Every major component has two instances in active and standby mode, backing up each other. There are two IP Peering Connections handled by two independent pairs of routers, two pairs of Cisco Unified Border Elements, and two audio conferencing bridges. If any of these components fails, its standby counterpart takes over. If the active peering link fails, the network will converge via the standby connection. All existing calls continue, but with a very brief interruption of the media flow. Cisco Unified Border Elements use the Out-of-Dialog OPTIONS ping mechanism to monitor the operational state of each other. Cisco Unified Border Elements at the customer site also monitor the Cisco Unified CM cluster using the Out-of-Dialog OPTIONS ping mechanism. Failure in responding to the ping results in removal of the unresponsive element from the dial-peer list of the sender, which commences routing all new calls via the standby instance. In case the active WebEx audio bridge fails, all calls associated with the bridge are terminated and the standby WebEx audio bridge is activated. WebEx will then prompt the users with a new number to connect to the newly activated bridge, which also re-dials all system-originated calls (callbacks) from before the failure.

Consider the following guidelines when deploying Cisco WebEx Cloud Connected Audio:

- Cisco recommends using Cisco Unified CM 8.5 or later release with the WebEx CCA deployment.
- Cisco recommends using a dedicated Cisco Unified Border Element for the WebEx CCA deployment to ensure a sound architecture and easy troubleshooting.
- Cisco Unified Border Element can be deployed on either a Cisco Integrated Services Router (ISR) or an Aggregated Services Router (ASR), depending on the audio port capacity requirements.
- Use an access control list (ACL) instead of packet inspection to restrict traffic in the firewall on the IP Peering link.
- The system administrator must provide at least one toll and one toll-free number for guest dial-in.
- If an audio codec other than G.711 is desired, use a transcoder to transcode the audio stream to G.711 before sending it to WebEx.
- One Direct Inward Dialing (DID) Digital Number Identification Service (DNIS) must be passed to the WebEx cloud via the Cisco Unified Border Element for all conferencing numbers.

For more information on Cisco WebEx Cloud Connected Audio, refer to the documentation available at

https://www.cisco.com/go/cwcca

Capacity Planning

For a given customer, the actual number of concurrent meetings is essentially unlimited. Different WebEx conferencing types have different capacities with respect to number of attendees. For a detailed product comparison table, refer to the *Cisco WebEx Web Conferencing Product Comparison*, available at

https://www.cisco.com/en/US/prod/ps10352/product_comparison.html

Network Traffic Planning

With the increased traffic out to the internet, it is important to consider network traffic planning. When planning for network traffic, the way that users use WebEx will make quite a bit of difference in the amount of traffic generated by the meeting. For example, if attendees use native presentation sharing (where the document is loaded to the WebEx site prior to sharing), it generates far less data than if they share their desktops. For a large enterprise, this can be important to understand to ensure correct traffic engineering, especially at the choke points in the network, such as the Internet access points. A preliminary estimate should be made around the average number of meetings to be hosted during the busy hour, along with the average number of attendees. Then, depending on the type and characteristics of these meetings, some projections on bandwidth requirements can be made. For more information regarding network traffic planning, please see the *Cisco WebEx Network Bandwidth* white paper, available at

https://www.cisco.com/c/en/us/products/collateral/conferencing/webex-meeting-center/white_paper_c11-691351.html

Design Considerations

Observe the following design considerations when implementing a Cisco WebEx SaaS solution:

- Collaborative meeting systems typically result in increased top-of-the-hour call processing loads. Cisco partners and employees have access to capacity planning tools with parameters specific to collaborative meetings to help calculate the capacity of the Cisco Unified Communications System for large configurations. Contact your Cisco partner or Cisco Systems Engineer (SE) for assistance with sizing of your system. For Cisco partners and employees, the Cisco Collaboration Sizing Tool is available at https://cucst.cloudapps.cisco.com/landing.
- All connections from WebEx clients are initiated out to the cloud. Typically, opening pinholes in network firewalls is not required as long as the firewalls allow intranet devices to initiate TCP connections to the Internet.
- Provision sufficient bandwidth for conference video and data traffic. See Network Traffic Planning, page 11-48, for details.
- Based upon business requirements, design decisions have to be made about the following:
  - User creation and authentication options (see User Profile, page 11-30, for details)
  - Meetings scheduling options (see Scheduling, page 11-30, for details)

- Cisco WebEx SaaS uses the multi-layer security model, and security extends from the WebEx infrastructure to the organization and individual meeting layer. There are various security options available, and depending on the business requirements., an organization can implement different levels of security. For security options and considerations, refer to the white paper *Unleash the Power of Highly Secure, Real-Time Collaboration*, available at

  https://www.cisco.com/en/US/products/ps12584/prod_white_papers_list.html

- For more details on the various Cisco Collaboration client offerings and how they fit into Cisco conferencing solutions, see the chapter on Collaboration Endpoints, page 8-1.

Cisco WebEx Meeting Center Video Conferencing

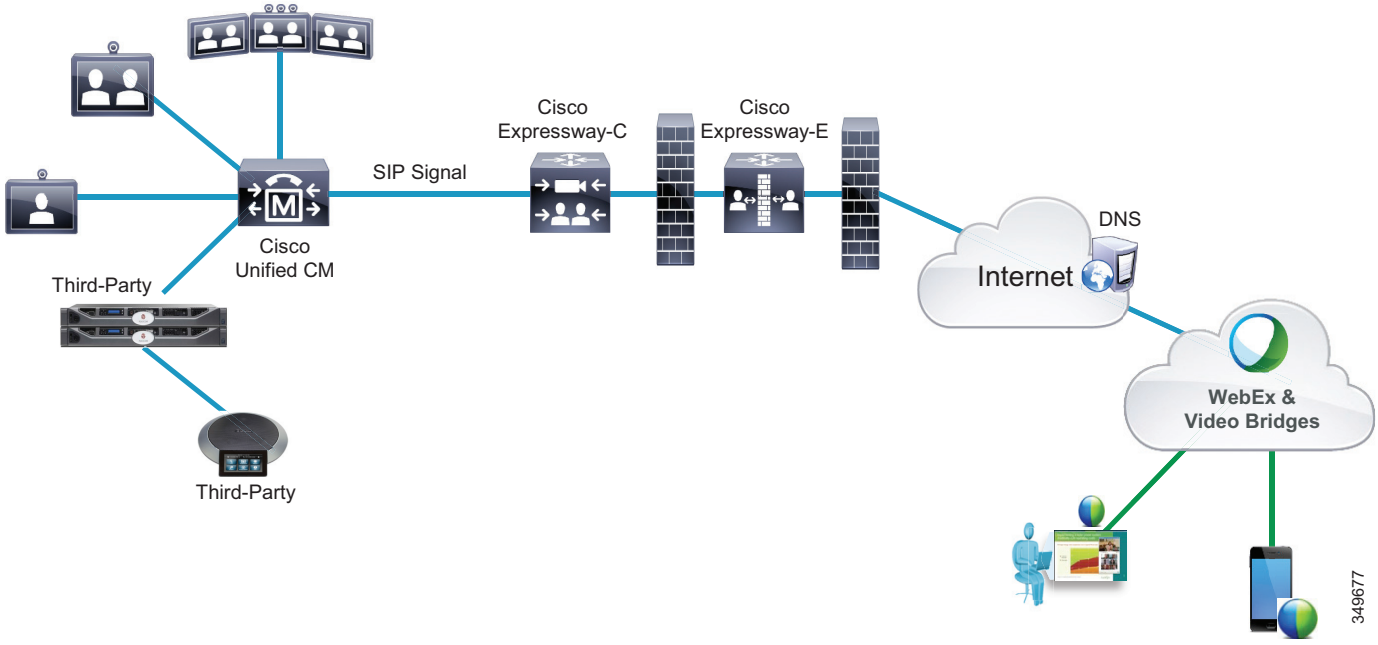
Cisco WebEx Meeting Center Video Conferencing provides a consistent, scalable virtual meeting room experience that combines business quality video, audio, and data sharing capabilities into a single solution delivered through Cisco WebEx Collaboration Cloud. Cisco WebEx Meeting Center Video Conferencing is included as part of the Cisco WebEx Meeting Center subscription purchase. It integrates with the Cisco Collaboration infrastructure and applications such as Cisco Unified CM and Cisco Expressway. Participants can join Cisco WebEx Meeting Center Video Conferencing meetings using WebEx clients, Cisco TelePresence, Cisco Jabber, or other third-party standards-based endpoints (SIP or H.323). It also provides a simple and highly secure collaboration solution from the Cisco WebEx Collaboration Cloud, and participants can join the meeting regardless of their location using any device of their choice (desktop, mobile, or video endpoint). With Cisco WebEx Meeting Center Video Conferencing, users can invite others to join their personalized, always-available meeting rooms anytime, or the meeting organizer can reserve the needed rooms and resources for scheduled meetings using the productivity tools.

Architecture

Figure 11-17 illustrates the Cisco WebEx Meeting Center Video Conferencing architecture using SIP video. This architecture consists of the enterprise collaboration network and the WebEx Collaboration Cloud where all the conferencing resources are hosted, and they are connected via the Internet. The enterprise collaboration network encompasses Cisco Unified Communications Manager (Unified CM) and Cisco Expressway, and Unified CM connects with Cisco Expressway-C over a SIP trunk. Cisco Unified CM provides the call routing and call control functions for the registered video devices. Cisco Expressway provides a secure firewall traversal mechanism for calls between the enterprise and WebEx Collaboration Cloud, and it routes the video calls to the WebEx Cloud via the DNS zone configured inside Cisco Expressway-E. In addition, Cisco Expressway provides mobile and remote access capability to the supported Cisco video endpoints so that they can register with Unified CM outside of the enterprise. In order for a participant to join the meeting and share content, the SIP device must support URI dialing and Binary Floor Control Protocol (BFCP). Without BFCP, content cannot be shared and will be seen embedded in the main video.

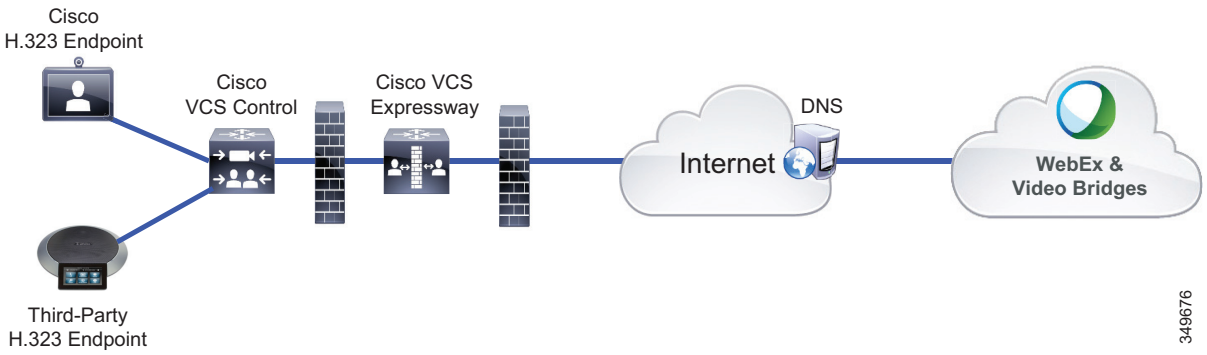
Note For existing Cisco VCS customers, using VCS Control as a SIP Registrar for SIP endpoints and VCS Expressway for firewall traversal is supported with the deployment.

Figure 11-17 Cisco WebEx Meeting Center Video Conferencing Architecture Using SIP Video



Cisco WebEx Meeting Center Video Conferencing architecture also support H.323 video devices (see Figure 11-18). In this architecture, Cisco VCS Control is the gatekeeper and provides call control for the registered H.323 endpoints. Cisco VCS Expressway provides a secure firewall traversal mechanism for calls between the enterprise and WebEx Cloud, and it routes the video calls to WebEx Cloud via the DNS zone configured inside Cisco VCS Expressway. In order for a participant to join the meeting and share content, the H.323 device must support Annex O for URI dialing and H.239 for content sharing. Without H.239, content cannot be shared and will be seen embedded in the video. In addition, H.323 devices must support either the H.245 User Input or RFC 2833 method of DTMF signaling in order to use interactive voice response (IVR) to start a meeting as a host or to join a meeting before the host joins.

Figure 11-18 Cisco WebEx Meeting Center Video Conferencing Architecture Using H.323 Video



Alternatively, Cisco WebEx Meeting Center Video Conferencing can be deployed using H.323 video without a call control system (see Figure 11-19). In this architecture, the H.323 device does not register to any gatekeeper; and when the user dials the URI, the call is routed using DNS through the firewall to the WebEx Cloud. Make sure the necessary ports on the firewall are opened so that signaling and media can pass through. For port range details, refer to the Collaboration Help article available at https://collaborationhelp.cisco.com/article/en-us/WBX264.

Figure 11-19 Cisco WebEx Meeting Center Video Conferencing Architecture Using H.323 Video Without a Call Control System

Staring with Cisco WebEx Meeting Center version WBS32.9, users can join meetings from H.323 video systems by dialing the regional IP address as shown in the meeting invite. However, if the user dials another region's IP address (IP address not shown in the meeting invite), the user will not be allowed to join the meeting.

Irrespective of SIP or H.323 devices used in the deployment, WebEx Cloud can perform the interworking between protocols. There are requirements for video devices to be used in a Cisco WebEx Meeting Center Video Conferencing deployment. For details, refer to the *Cisco WebEx Meeting Center Enterprise Deployment Guide for Video Device-Enabled Meetings,* available at

https://collaborationhelp.cisco.com/article/en-us/nmdp0hq

For each participant on a video device, the audio, video, and content sharing are sent over the IP connection to WebEx Cloud, where the media are mixed with other participants, and the mixed audio, active speaker video, and content sharing are sent back to the device for display.

Cisco WebEx Meeting Center Video Conferencing uses H.264 video for active speaker and content sharing. Depending on the capability of the device and the bandwidth available, Cisco WebEx Meeting Center Video Conferencing supports active speaker video up to 720p at 30 frames per second (fps) and content video up to 720p on video devices as well as WebEx clients. The WebEx meeting client has a video floor of 180p for active speaker video at a minimum bit rate of 1.2 Mbps. If the minimum bit rate cannot be maintained due to network conditions (severe packets loss, for example), the WebEx client will stop receiving the active speaker video, but it still receives content sharing as well as conference audio and sends its video to other participants. The WebEx client will periodically perform bandwidth retest and automatically reestablish active speaker video when network conditions stabilize. During the meeting, WebEx allocates the bandwidth based upon the least capable device among all WebEx clients in the conference (excluding devices running below the video floor), with a maximum bandwidth of 4 Mbps. However, if the least capable device leaves the conference, the bandwidth will be reallocated based upon the next least capable device that runs the WebEx meeting client. The allocated bandwidth determines the resolution used to display the video on the WebEx clients.

Each Cisco WebEx Meeting Center Video Conferencing session has an associated video address URI and URL. Participants dial the URI or receive callback on the video device or click on the URL to bring up the WebEx meeting client to join the meeting. A Cisco WebEx Meeting Center Video Conferencing meeting can be one of the following types:

Scheduled meeting

Users can use WebEx Productivity Tools (PT) to schedule Cisco WebEx Meeting Center Video Conferencing meetings. Productivity Tools is a suite of tools, including an Outlook plug-in, that allows users to schedule meetings quickly and easily within the email client. This tool suite provides seamless integration with the user's calendar, and users can schedule meetings and send the invitations to all participants directly inside the email client with a single transaction. Alternatively, users can schedule Cisco WebEx Meeting Center Video Conferencing meetings from the WebEx portal, but the host must first schedule the meeting from WebEx and then create an invitation with meeting details attached and send it to all the participants.

Using Cisco TMS 15.2 and TMSXE 5.2, WebEx Productivity Tools can be utilized to schedule Cisco WebEx Meeting Center Video Conferencing meetings with One Button to Push (OBTP). Internally, Cisco TMS creates an externally hosted conference using the SIP URI to Cisco WebEx Meeting Center Video Conferencing as the dial string. Also, Cisco TMS must have the default conference type set to OBTP.

Note Using Cisco TMS and TMSXE for Cisco WebEx Meeting Center Video Conferencing OBTP does not require integration with the on-premises video conferencing infrastructure. On the other hand, if TMS and TMSXE are integrated with the on-premises video conferencing infrastructure, they can be used for Cisco WebEx Meeting Center Video Conferencing OBTP at the same time.

Beginning with Cisco WebEx Meeting Center version WBS32.6, users can schedule and start meetings on their WebEx Personal Room directly from their Google Calendar. In order to use this feature, users need to install Cisco WebEx Scheduler extension from the Google Web Store, and the site administrator needs to enable the Google Calendar option for the WebEx site. When a user signs into the Cisco WebEx Scheduler extension with his WebEx account, this allows the user's Google account to use the WebEx service. Only one WebEx account can be linked to the user's Google account. For details on Cisco WebEx Scheduler for Google Calendar configuration and limitations, refer to the following articles:

- https://collaborationhelp.cisco.com/article/en-us/j5bm2v
- https://collaborationhelp.cisco.com/article/en-us/9pq6jc

Permanent meeting

Meetings can be hosted in the user's personal room. Personal rooms can be enabled at the site level or per-user level in the WebEx site. When personal rooms are enabled, a fixed URI and URL are assigned to each user, and participants can use them to join the user's personal room. This personal room belongs to the designated user and is always on. Thus, the user can use his room for his meetings and can send an invitation to all participants with his room's URI and URL attached. With Cisco Spark Calendar Service, users can add @webex to the location field of an Outlook calendar invitation, and Calendar Connector will automatically populate the invitation with the user's personal room information. See the chapter on Mobile Collaboration, page 21-1, for more details.

Instant meeting

A user can create an instant meeting from the WebEx portal or by using the WebEx Productivity Tools, and the meeting will start immediately. Using the Meet Now configuration option, the instant meeting can be initiated from the Meeting Center, the user's personal room, or Cisco Jabber Desktop.

Security

Cisco WebEx Meeting Center Video Conferencing supports encrypted signaling and media, or a combination of encrypted and non-secure signaling and media, between the enterprise network and WebEx Cloud. For end-to-end encryption, customers can turn on encrypted signaling and media in the enterprise and use encrypted signaling and media between the enterprise network and the WebEx Cloud. A certificate has to be uploaded to Cisco Expressway-E to ensure that proper handshaking takes place for encrypted signaling to be functional. That certificate can be either self-signed or signed by a trusted Root Certificate Authority (CA). For more information, refer to the latest version of the *Cisco WebEx Meeting Center Video Conferencing Enterprise Deployment Guide,* available at

https://www.cisco.com/c/en/us/support/conferencing/webex-meeting-center/products-installation-and-configuration-guides-list.html

For SIP-based calls, Cisco WebEx Meeting Center Video Conferencing supports four levels of security (in order of preference):

- Encrypted TLS signaling with CA-signed certificates and SRTP media encryption
- Encrypted TLS signaling with self-signed certificates and SRTP media encryption
- Non-secure TCP signaling with SRTP media encryption
- Non-secure TCP signaling with non-secure RTP media

Make sure to open the network ports on the firewall so that inbound and outbound traffic for signaling and media can pass through. For port range details, refer the WebEx article available at https://collaborationhelp.cisco.com/article/en-us/WBX264.

All Cisco WebEx Meeting Center Video Conferencing meetings require the presence of the host to start the meeting. If the guests join before the host, they will be in the waiting room and cannot talk to each other until the host joins. In addition, a host PIN is required when the host joins the meeting from a video device.

Inside the user's personal meeting room, a Lock Room button is available that can be used to lock the room and prevent other participants from entering the user's personal room. When the room is locked and a participant tries to enter the room, that participant will be blocked until the host admits him or unlocks the room. This button is useful in case a user's personal room is used for back-to-back meetings and the host has not finished with the first meeting. The host can lock the room to prevent participants of the second meeting from entering until he finishes with the first meeting and unlocks the room.

Audio Deployment Options

For Cisco WebEx Meeting Center Video Conferencing participants using video devices, their audio, video, and content sharing are sent and received over the IP connection between the WebEx Cloud and the video devices. For WebEx client participants, Cisco WebEx Meeting Center Video Conferencing supports all audio options available for the classic WebEx Meeting Center, which include:

- WebEx Cloud Connected Audio
- WebEx Audio using VoIP
- WebEx Audio using PSTN
- Teleconferencing service provider audio

High Availability

In the enterprise collaboration network, utilize the clustering option with Cisco Unified CM and Cisco Expressway to provide redundancy for call control with video devices and firewall traversal calls. If the primary server fails, the backup server can take over the call control and call handling functions.

For Cisco Unified CM clustering, see the chapter on Call Processing, page 9-1.

For Cisco Expressway clustering, refer to the latest version of the *Cisco Expressway Cluster Creation and Maintenance Deployment Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html

Capacity Planning

Cisco WebEx Meeting Center Video Conferencing meetings support up to 25 standards-based video devices, 500 WebEx participants with video enabled, and 500 WebEx participants with audio only.

Note Each screen in a multi-screen video device counts as one video device. For example, if a triple-screen immersive system joins the Cisco WebEx Meeting Center Video Conferencing meeting, it consumes 3 video devices from the video device capacity limit.

Capacity planning for Cisco WebEx Meeting Center Video Conferencing involves sizing of the components running within the enterprise. The components could include:

- Cisco Unified CM

  Ensure that Unified CM has enough resources and capacity to handle the traffic generated by the video endpoints and IP phones for Cisco WebEx Meeting Center Video Conferencing meetings. For capacity details, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

- Cisco Expressway

  Cisco Expressway must provide enough resources to handle the traversal call traffic for the deployment. For capacity details, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Network Traffic Planning

Network traffic planning for Cisco WebEx Meeting Center Video Conferencing consists of the following elements:

- WebEx Clients bandwidth

  The WebEx meeting client uses Scalable Video Coding (SVC) technology to send and receive video. It uses multi-layer frames to send video, and the receiving client automatically selects the best possible resolution to receive video that typically requires 1.2 to 3 Mbps of available bandwidth. For more information regarding network traffic planning for WebEx clients, refer to the *Cisco WebEx Network Bandwidth* white paper, available at

  https://www.cisco.com/c/en/us/products/collateral/conferencing/webex-meeting-center/white_paper_c11-691351.html

- Bandwidth for video device from enterprise to WebEx Cloud

  For optimal SIP audio and video quality, Cisco recommends setting up the video bandwidth for at least 1.5 Mbps per device screen in the region associated with the endpoint registering with Cisco Unified CM. For example, if a triple-screen device registers with Unified CM, video bandwidth of 4.5 Mbps should be allocated in the associated region.

Design Considerations

Consider the following recommendations when deploying Cisco WebEx Meeting Center Video Conferencing:

- Enable UDP for media streaming in the firewalls for the optimal video experience.
- Open network ports on firewalls to allow inbound and outbound signaling and media traffic. For details, refer to the latest version of the *Cisco WebEx Meeting Center Video Conferencing Enterprise Deployment Guide,* available at

  https://www.cisco.com/c/en/us/support/conferencing/webex-meeting-center/products-installation-and-configuration-guides-list.html

- Ensure that Binary Floor Control Protocol (BFCP) is enabled in the Unified CM Neighbor Zone in Cisco Expressway-C and that BFCP is also enabled in the SIP profile associated with the SIP trunk between Unified CM and Expressway-C.
- For information on the video devices tested with Cisco WebEx Meeting Center Video Conferencing, refer to the *WebEx Video Compatibility and Support* article available at

  https://collaborationhelp.cisco.com/article/en-us/ipxxr2

Cisco WebEx Meetings Server

Cisco WebEx Meetings Server is a highly secure, fully virtualized, private cloud conferencing solution that combines audio, video, and web conferencing in a single solution. Cisco WebEx Meetings Server addresses the needs of today's companies by presenting a comprehensive conferencing solution with all the tools needed for increased employee productivity as well as support for more dynamic collaboration and flexible work styles. Existing customers can build on their investment in Cisco Unified Communications and extend their existing implementation of Cisco Unified Communications Manager to include conferencing using the SIP architecture. In addition, Cisco WebEx Meetings Server leverages many capabilities from Cisco Unified CM to perform its functions; for example:

- Use the SIP trunk connection with Unified CM to conduct teleconferencing
- Utilize Unified CM's SIP trunk secure connection support for secure conferencing
- Integrate with legacy or third-party PBXs through Unified CM
- Leverage Unified CM's dual stack (IPv4 and IPv6) capability to support IPv6

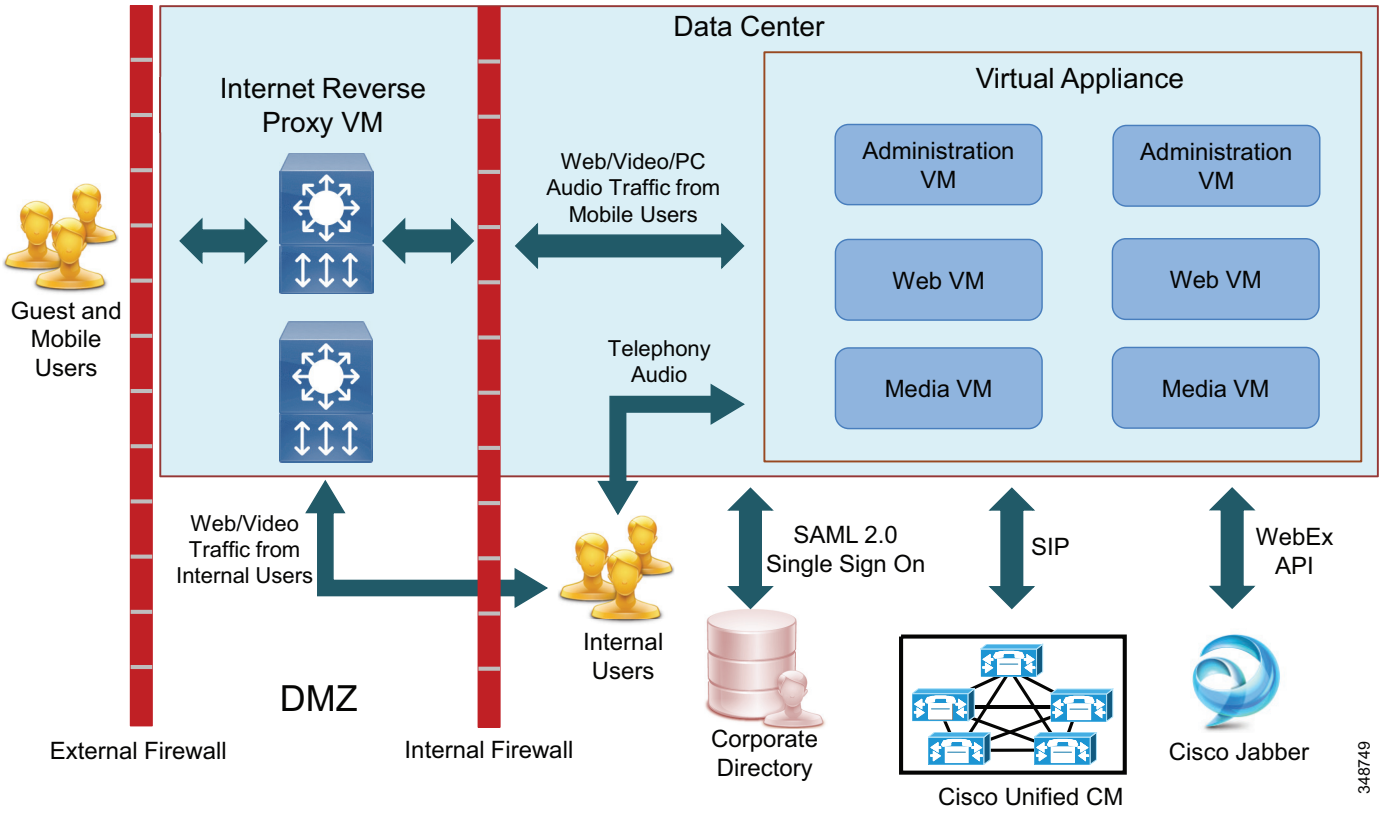
These capabilities are discussed in more detail in the following sections.

Architecture

Cisco WebEx Meetings Server is a fully virtualized, software-based solution that runs on Cisco Unified Computing System (UCS). It uses the virtual appliance technology for rapid deployment of services. Virtual appliance simplifies the task of managing the system. For example, using the hypervisor technology, system components can easily be moved around for maintenance, or system components can easily be rolled back to a working version if problem arises. The virtual appliance is distributed in the form of an industry standard format, Open Virtual Appliance (OVA). All the software components required to install WebEx Meetings Server are packaged inside the OVA. Traditionally, using an executable installer to install individual software components would take hours to deploy the software. However, using OVA can significantly reduce the amount of time required to deploy the software because all software components are pre-packaged inside the file. Thus, virtual appliance technology can help tremendously to reduce the deployment time for Cisco WebEx Meetings Server.

Figure 11-20 shows the high-level architecture for Cisco WebEx Meetings Server using the non-split horizon network topology. (For details on the non-split horizon network topologies, refer to the *Cisco WebEx Meetings Server Planning Guide*, available at https://www.cisco.com/en/US/products/ps12732/products_installation_and_configuration_guides_list.html.) Inside the virtual appliance, there could be one or more virtual machines (VMs) running. These are the administration, web, and media virtual machines. The administration and web virtual machines serve as the back-end processing for the administration and WebEx sites. These sites handle tasks that happen before and after the meeting, such as configuration, scheduling/joining meetings, and recording playback. The media virtual machine provides resource allocation, teleconference call control, and media processing (voice, video, and data) during the meeting. The number of virtual machines running inside the virtual appliance depends on the capacity desired and on whether high availability is needed. This provides various options for deployment size.

Figure 11-20 Cisco WebEx Meetings Server High-Level Architecture

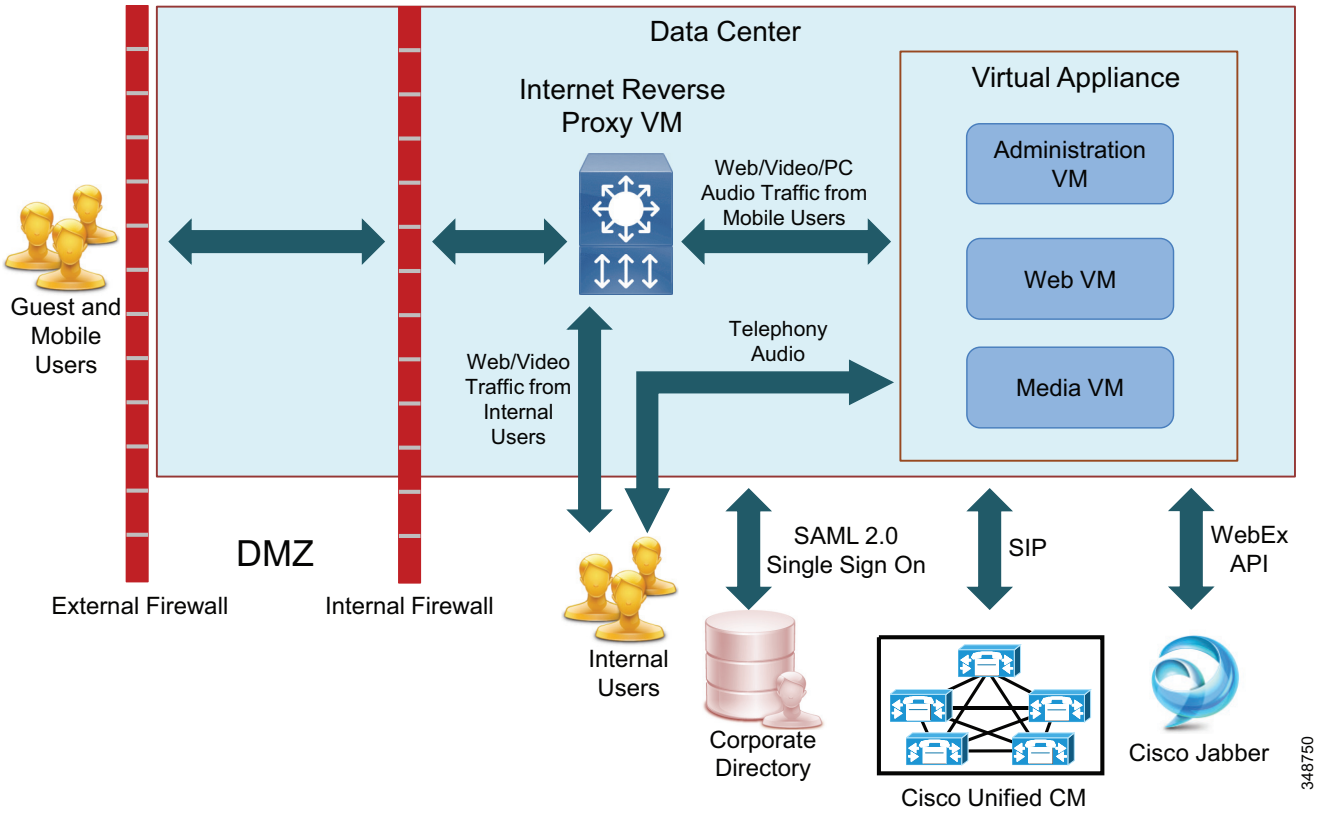


Cisco WebEx Meetings Server offers the option of deploying the Internet Reverse Proxy (or edge servers) in the DMZ to facilitate external access. This option provides two advantages. First, all external participants can securely access the WebEx conferences from the internet without going through a VPN. Second, mobile users can join the meetings from a mobile device anywhere as long as there is internet connectivity. Note that the Internet Reverse Proxy is mandatory if mobile client access is enabled.

Internet Reverse Proxy is used to terminate all inbound traffic from the internet inside the DMZ. The content is then forwarded to the internal virtual machines through an encrypted Secure Socket Layer (SSL) or Transport Layer Security (TLS) tunnel. This encrypted tunnel is established by the internal virtual machines connecting outbound to the Internet Reverse Proxy. Therefore, there is no need to open TCP ports inbound from the DMZ to the internal network on the internal firewall. However, some outbound ports from the internal network need to be opened on the internal firewall to allow communication with the Internet Reverse Proxy in the DMZ.

All end-user sessions are 100% encrypted using industry standard Secure Socket Layer (SSL) and Transport Layer Security (TLS). All traffic between the virtual machines is sent over the secure channel. Federal Information Processing Standard (FIPS) encryption can also be turned on by a single policy setting, providing US Department of Defense (DoD) level security. Alternatively, the Internet Reverse Proxy can be deployed behind the internal firewall as shown in Figure 11-21.

Figure 11-21 Internet Reverse Proxy Behind the Internal Firewall



For security concerns, an organization would typically take several months to get approval in deploying a component inside the DMZ. Using this methodology, it could eliminate any DMZ components and bypass the approval process to get the WebEx Meetings Server deployment done quickly. All internet traffic (HTTP on port 80 and SSL on port 443) to the external firewall should be forwarded to the internal firewall. This will minimize the number of ports that need to be opened in the external and internal firewalls. However, placing the Internet Reverse Proxy inside the internal network implies that inbound internet traffic will terminate in the internal network. Although direct internet access to the internal network could be controlled by the firewalls, not all organizations allow terminating internet traffic directly on their internal network. Ensure that this deployment does not violate your organization's IT policy before choosing this option.

In a large enterprise deployment, an organization would require the Single Sign On (SSO) capability to allow end users to sign in using their corporate credentials. Cisco WebEx Meetings Server can connect to the corporate LDAP directory using the industry standard SAML 2.0 for SSO.

Note Cisco WebEx Meetings Server supports Meeting Center only.

Note Starting with Cisco WebEx Meetings Server 1.1, Cisco Jabber integrated with the Cisco Unified CM IM and Presence Service can be used to join or start meetings hosted on WebEx Meetings Server. For Cisco Jabber support details, refer to the *Cisco WebEx Meetings Server System Requirements*, available at https://www.cisco.com/en/US/products/ps12732/prod_installation_guides_list.html.

Cisco Unified CM Integration

Cisco WebEx Meetings Server support both Cisco Unified CM and Session Management Edition (SME). Cisco Unified CM is a central piece of the WebEx Meetings Server architecture that allows the following:

- Attendees joining the teleconference by means of Cisco IP Phone or PSTN
- Integration of legacy or third-party PBXs with Cisco WebEx Meetings Server

Cisco Unified CM integrates with WebEx Meetings Server by means of SIP trunks to provide inbound and callback call control. Customer can choose to turn on security and run Transport Layer Security (TLS) and Secured Real-time Transport Protocol (SRTP) over the SIP trunk connection. A SIP trunk is configured in Unified CM with a destination address of the Load Balancer in WebEx Meetings Server, and then a route pattern (match the call-in access number configured in WebEx Meetings Server) must be used to route calls via the SIP trunk. A second SIP trunk is configured in Unified CM with a destination address of the Application Server in WebEx Meetings Server, and then a SIP route pattern must be used to route calls via the SIP trunk. When an attendee dials the access number to join the meeting, the first SIP trunk is used to send the call. After the call is connected and the caller enters the meeting ID, the Load Balancer issues a SIP REFER to Unified CM to send the caller to the Application Server that hosts the meeting via the second SIP trunk.

The system administrator can configure a SIP trunk in WebEx Meetings Server that points to a Unified CM to perform callback. Attendees can provide a callback number and have the system out-dial the number to the attendees to join the bridge. In the case of attendees requesting callback, the WebEx Meetings Server sends the SIP request to Unified CM along with the callback number via the configured SIP trunk. It is imperative for Unified CM to be able to resolve all dial strings received from a callback request to join the meetings. Callbacks may also be disabled system-wide by means of site administration settings. Unified CM is in control of all toll restrictions to various countries or other numbers that most enterprises will block, because WebEx Meetings Server does not have any toll restriction blocking itself.

WebEx Meetings Server supports the bidirectional SIP OPTIONS ping mechanism. The ping response from the remote end indicates that the remote end is active and whether it is ready to accept calls. Based on the response, WebEx Meetings Server or Unified CM can determine whether to send calls on the current SIP trunk or look for an alternate SIP trunk (if configured) to send calls. Note that SIP OPTIONS ping is supported in Cisco Unified CM 8.5 and later releases. Due to this reason, Cisco recommends using a compatible Cisco Unified CM version that supports SIP OPTIONS ping for Cisco WebEx Meetings Server deployment. For the list of compatible Unified CM versions, refer to the compatibility matrix in the *Cisco WebEx Meetings Server System Requirements*, available at

https://www.cisco.com/en/US/products/ps12732/prod_installation_guides_list.html

Note Cisco WebEx Meetings Server supports SIP trunk connection with Cisco Unified CM only.

Legacy PBX Integration

Some organizations that have a legacy PBX and are not ready to fully migrate to a Cisco Unified Communications solution, might want to use Cisco WebEx Meetings Server with their system for conferencing. Cisco Unified CM can be used to bridge the legacy PBX and Cisco WebEx Meetings Server together. Cisco WebEx Meetings Server can see only Unified CM and does not even know the PBX is behind Unified CM. As long as Unified CM can interoperate with the organization's PBX, Cisco WebEx Meetings Server can integrate with the organization's PBX. This integration can provide several benefits:

- Allow users in the legacy system to experience the new technology
- Allow an organization to adopt the new technology gradually, at its own pace
- Protect the customer's investment in existing technology while allowing them to migrate to Cisco technology gradually

For further details on PBX interoperability with Unified CM, refer to the documentation available at

https://www.cisco.com/en/US/solutions/ns340/ns414/ns728/networking_solutions_products_genericcontent0900aecd805b561d.html

IPv6 Support

Cisco WebEx Meetings Server supports IPv4 only or dual stack (IPv4 and IPv6) addressing for telephony audio, while telephony signaling remains at IPv4. Audio streams can be IPv4, IPv6, or a mix of IPv4 and IPv6 in the same meeting. Cisco WebEx Meetings Server supports Alternate Network Address Types (ANAT) to enable both IPv4 and IPv6 media addressing in the Session Description Protocol (SDP) during the SIP Offer and Answer exchange on the SIP trunk with Unified CM to establish a media connection using the preferred addressing scheme.

Both IPv4 and IPv6 devices can be used for teleconferencing. With IPv6 devices, Cisco WebEx Meetings Server leverages Unified CM's capacity to translate the IPv6 signaling to IPv4 and transport it over a SIP trunk to the Cisco WebEx Meetings Server. With the telephony media addressing, Cisco WebEx Meetings Server can convert between IPv4 and IPv6. Therefore, Cisco WebEx Meetings Server can support IPv6 without any expensive MTP resources.

With ANAT, Cisco WebEx Meetings Server can support IPv6 telephony audio without the support of IPv6 telephony signaling. However, ANAT must be supported on both ends of the Unified CM SIP trunk. Be sure to enable ANAT on the Unified CM SIP trunk, otherwise there will be a failure to establish the call when attendees request callback or attempt to dial in.

If the WebEx Meetings Server has IPv6 enabled, ANAT headers will be included in the media offer. WebEx Meetings Server will always answer with ANAT headers if the media offer includes ANAT headers. The following paragraphs describe the media address version selection process between the IPv6-enabled WebEx Meetings Server and the dual-stack Unified CM using the ANAT header.

When WebEx Meetings Server sends a call to Unified CM, the SDP offer contains both IPv4 and IPv6 media addresses. If the called device is IPv6, Unified CM chooses IPv6 for the media connection and answers with the IPv6 media address in the SDP; if the called device is dual-stack, Unified CM uses the **IP Addressing Mode Preference for Media** parameter to determine the address version in the answer SDP. If the parameter is set to IPv6, then IPv6 will be used for the media connection.

When Unified CM sends a call to the WebEx Meetings Server through the SIP trunk, WebEx Meetings Server receives the SDP offer with an ANAT header. If the SDP offer contains both IPv6 and IPv4 media addresses, WebEx Meetings Server answers with the higher precedence address version specified in the ANAT header, which would be IPv6 in this case. If the SDP contains only an IPv6 address, WebEx Meeting Server answers with an IPv6 media address.

For information on deploying IPv6 in a Cisco Unified Communications system, refer to the latest version of *Deploying IPv6 in Unified Communications Networks with Cisco Unified Communication Manager*, available at

https://www.cisco.com/go/ucsrnd

High Availability

Cisco WebEx Meetings Server uses the N+1 redundancy scheme to ensure system availability in the event of component failures. High availability is achieved by adding a local, redundant system to the primary system within the same data center. At the system level, virtual machines and components inside run in active/active mode. If one component goes down, the system restarts the component. Status information is exchanged between system components. Using this status information, the system is able to distribute the requests evenly among the active components. Depending on the deployment size, the number of virtual machines in the backup or redundant system might or might not be the same as in the primary system.

In the high availability system, when the virtual machine hosting the meeting goes down, affected meeting clients will automatically reconnect to the available service within a short period of time. However, depending on the nature of the failure and which component has failure, not all clients and meetings would be affected. For descriptions of high availability system behavior after a component failure, refer to the latest version of the *Cisco WebEx Meetings Server Administration Guide*, available at

https://www.cisco.com/c/en/us/support/conferencing/webex-meetings-server/products-installation-guides-list.html

Virtual IP Address

Inside the high availability system, there is a second network interface in the active administration and Internet Reverse Proxy virtual machine that is configured with the virtual IP address. The administration and WebEx site URLs use this virtual IP address to access the administration and WebEx sites. In the event of failover, the virtual IP address is moved over to the new active virtual machine. Thus, it provides access redundancy to the administration and WebEx site.

Multiple Data Center Design

Cisco WebEx Meetings Server can be deployed in multiple data centers (up to maximum of 2) for geographic redundancy or disaster recovery. In this deployment, there are two WebEx Meetings Server systems with identical deployment size, one in each data center, that are joined together to form a single logical system running in active/active mode. The first system added to the multi-data center system is the primary, and the system that is added after that is the secondary. When the secondary system is added to the multi-data center system, all its global data are overwritten with the data from the primary system and only configuration parameters local to the data center are preserved. Refer to the *Cisco WebEx Meetings Server Administration Guide* for details on the types of data that will be overwritten and preserved. Within each data center, there are local Unified CM instances for handling teleconferencing. System status is exchanged, and information about users and meetings is synchronized across data center

peers over an encrypted SSL link. Administrators use a single URL to manage the systems, and participants use a single URL or one set of dial-in numbers to join the meeting. When participant join a meeting via the client, the system automatically chooses the data center closest to the participant to host the meeting, and the meeting is cascaded across data centers.

In the event of failure, if one component goes down in the data center, the system restarts that component. If the whole data center goes down, the surviving data center takes over without any manual intervention, and the system still runs with full capacity. When this happens, affected meeting clients automatically reconnect to the service in the surviving data center within a short period of time. However, depending on the nature of the failure and state of the client, the recovery mechanism might be different and would follow the same behavior as the high availability system. For detail descriptions, refer to the latest version of the *Cisco WebEx Meetings Server Administration Guide*, available at

https://www.cisco.com/c/en/us/support/conferencing/webex-meetings-server/products-installation-guides-list.html

Consider the following information when using the multiple data center design:

- Configure NTP in all data centers.
- A multi-data center license is required for the WebEx Meetings Server system in each data center. Install the licenses onto the primary data center system before joining the data centers.
- A deployment size of 50 users per system is not supported, but larger system sizes are supported.
- Running a high availability system within the data center is not supported.
- Deploy local Unified CM instances in each data center.
- Joining the systems together will not increase the total system capacity.
- Either both data centers or neither data center can have Internet Reverse Proxy deployed.

Capacity Planning

The capacity of WebEx Meetings Server depends on the platform of choice and the number of conferencing nodes running in the deployment. For capacity planning details, see the section on Collaborative Conferencing, page 25-44.

Storage Planning

If recording meetings is a requirement, sufficient disk space should be allocated on the Network Attached Storage (NAS) device to store the recordings. For disk space allocation detail, refer to the *Meeting Recordings* section in the *Cisco WebEx Meetings Server Planning Guide*, available at

https://www.cisco.com/en/US/products/ps12732/products_installation_and_configuration_guides_list.html

Network Traffic Planning

Network traffic planning for WebEx Meetings Server collaboration consists of the following elements:

- Call control bandwidth

  Call control bandwidth is extremely small but critical. Co-locating the WebEx Meetings Server with Unified CM helps protect against issues with call control. Remote locations need proper QoS provisioning to ensure reliable operation. Call control bandwidth is used for establishment of calls between WebEx Meetings Server and Unified CM, and the amount of bandwidth required for each call depends on how the attendees join the meeting. For an attendee dialing into the meeting, the call consumes approximately the same amount of bandwidth as making two SIP calls. For an attendee requesting callback, the call consumes approximately the same amount of bandwidth as making one SIP call. For details about call control bandwidth estimation for SIP calls and QoS provisioning, see the chapter on Network Infrastructure, page 3-1.

- Real-Time Transport Protocol (RTP) traffic bandwidth

  RTP traffic consists of voice and video traffic. Voice bandwidth calculations depend on the audio codec used by each device. (See the chapter on Network Infrastructure, page 3-1, for bandwidth consumption by codec type.) Video bandwidth can be calculated the same way as WebEx SaaS. (See Network Traffic Planning, page 11-33.)

- Web collaboration bandwidth

  Web collaboration bandwidth for WebEx Meetings Server can be estimated the same way as WebEx SasS. (See Network Traffic Planning, page 11-33.)

- Multiple data center deployment

  For proper operation and optimal user experience with this deployment, there are network requirements for maximum round-trip delay time (RTT) and minimum guaranteed bandwidth plus additional bandwidth for each cascaded meeting between data centers. For network requirement details, refer to the latest *Cisco WebEx Meetings Server Planning Guide and System Requirements*, available at

  https://www.cisco.com/en/US/products/ps12732/products_installation_and_configuration_guides_list.html

Design Consideration

The following additional design considerations apply to WebEx Meetings Server deployments:

- For scenarios where any WebEx Meetings Server components are separated by network firewalls, it is imperative to ensure the correct pinholes are opened for all required traffic.
- Collaborative meeting systems typically result in increased top-of-the-hour call processing load. Capacity planning tools with specific parameters for WebEx Meetings Server are available to Cisco partners and employees to help calculate the capacity of the Cisco Unified Communications System for large configurations. Contact your Cisco partner or Cisco Systems Engineer (SE) for assistance with sizing of your system. For Cisco partners and employees, the Cisco Collaboration Sizing Tool is available at https://cucst.cloudapps.cisco.com/landing.
- Using Transport Layer Security (TLS) and Secured Real-time Transport Protocol (SRTP) have no effect to the WebEx Meetings Server capacity. However, using TLS and SRTP does have an impact on Cisco Unified CM capacity.
- WebEx Meetings Server has no built-in line echo cancellation. Use an external device such as a Cisco Integrated Service Router (ISR) to provide echo cancellation functionality.

- For more details on the various Cisco Collaboration client offerings and how they fit into Cisco conferencing solutions, see the chapter on Collaboration Endpoints, page 8-1.
- Call admission control with WebEx Meetings Server is performed by Unified CM. With locations-based call admission control, Unified CM can control bandwidth to the WebEx Meetings Server system by placing the SIP trunk specific to WebEx Meetings Server in a location with a set amount of audio bandwidth allowed. Alternatively, Unified CM supports the use of Resource Reservation Protocol (RSVP), which can also provide call admission control. For further information regarding call admission control strategies, see the chapter on Bandwidth Management, page 13-1.
- Cisco recommends marking both the audio streams and video streams from WebEx Meetings Server as AF41 (DSCP 0x22) to preserve lip-sync. These values are configurable in WebEx Meetings Server Administration.
- Web conferencing traffic is encrypted in SSL and is always marked best-effort (DSCP 0x00).

Reference Document

For network requirements, network topology, deployment size options, and other deployment requirements and options for WebEx Meetings Server, refer to the *Cisco WebEx Meetings Server Planning Guide*, available at

https://www.cisco.com/en/US/products/ps12732/products_installation_and_configuration_guides_list.html

Cisco Collaboration Meeting Rooms Hybrid

Cisco Collaboration Meeting Rooms (CMR) Hybrid is a collaboration conferencing platform that combines the video experience of Cisco TelePresence Conferencing with the presentation experience of Cisco WebEx Meeting into a single meeting. Cisco WebEx and TelePresence are optimized to work with standards-based video endpoints and WebEx meeting clients. They help customers to extend the reach of the meetings and simplify the experience for all participants. Attendees on TelePresence endpoints and WebEx meeting clients can securely share two-way video, audio, and content among themselves. This platform brings together the user experiences from two conferencing systems and extends the collaboration to more users on more devices in more locations.

Cisco CMR Hybrid allows an organizer to schedule meetings using the familiar interface of Microsoft Outlook enabled by the WebEx Productivity Tools or with the Cisco TelePresence Management Suite (TMS). The host selects the participants, adds the preferred endpoints and the WebEx information, and sends the invitation to all attendees. Using the productivity tools, the attendees receive one meeting invitation with all the information about how to join through TelePresence or WebEx. The meetings can be launched using One Button To Push (OBTP) from the TelePresence endpoint, or Cisco TMS can automatically connect the endpoints with the meetings at the scheduled start time.

Architecture

As shown in Figure 11-22, the high-level architecture of Cisco CMR Hybrid consists of the enterprise collaboration network and the WebEx Cloud infrastructure that are connected through an IP connection. The enterprise collaboration network consists of Cisco Unified Communications Manager (Unified CM), Cisco Expressway-C and Expressway-E, TelePresence Bridge pools that are managed by TelePresence Conductor, and Cisco TelePresence Management Suite (TMS). Cisco Unified CM is the call processing platform that provides call routing and call control for the TelePresence endpoints within the enterprise. Cisco Expressway-C and Expressway-E route calls between the enterprise network and WebEx Cloud. Cisco Unified CM connects with Cisco Expressway-C and Cisco TelePresence Conductor over separate Best Effort Early Offer SIP trunks.

For details on integrating Cisco Unified CM with Cisco Expressway, refer to the latest version of the *Cisco Expressway and CUCM via SIP Trunk Deployment Guide*, available at

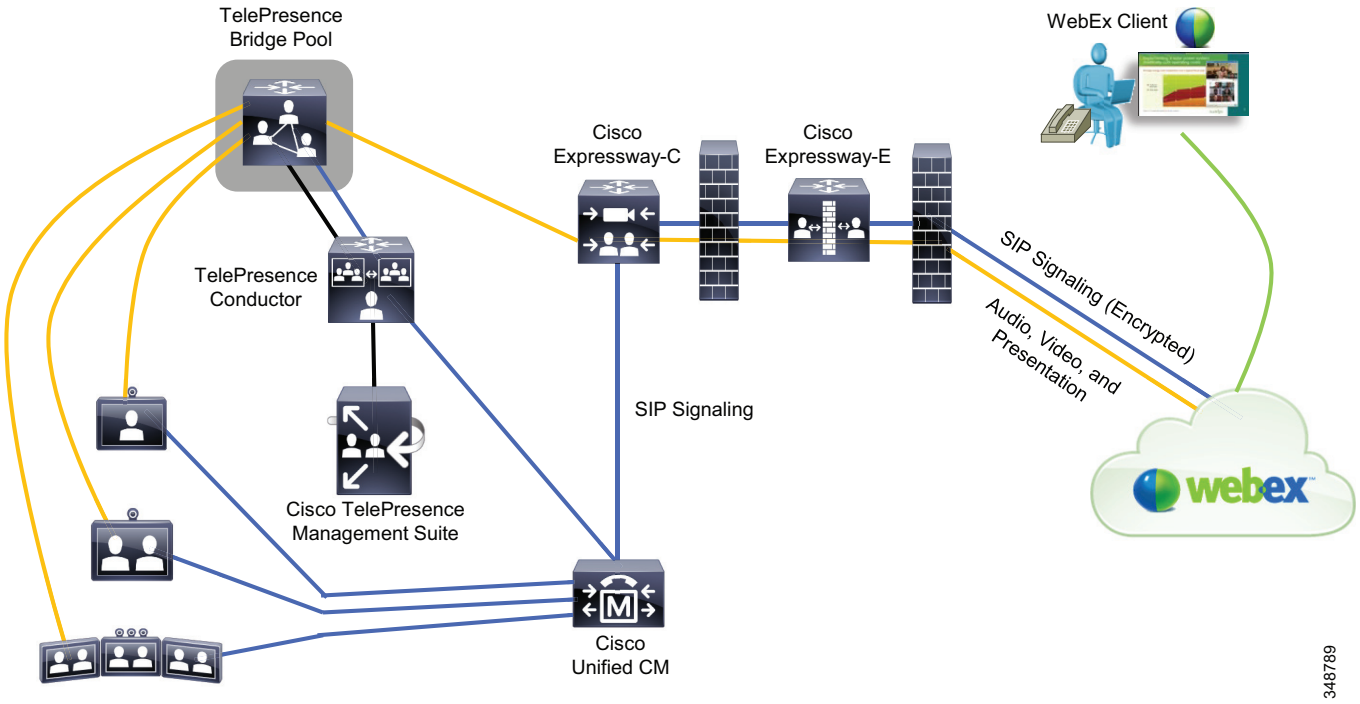
https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html

Note For existing Cisco VCS customers, deployment using Cisco VCS Control and Expressway in place of Cisco Expressway-C and Expressway-E is supported.

Note Deployment using a Best Effort Early Offer SIP trunk between Unified CM and the TelePresence Bridge without TelePresence Conductor is supported, but using TelePresence Conductor is recommended.

Cisco TelePresence Conductor selects a TelePresence Bridge from the pool to host the TelePresence conference. The TelePresence Bridge mixes the audio from the TelePresence endpoint participants and sends the mixed audio, the active speaker video, and the content sharing video to the WebEx Cloud using SIP. Similarly, the TelePresence Bridge receives the media (mixed audio, active speaker, and content sharing video) from the WebEx Cloud, cascades the audio into the TelePresence conference, and sends the content sharing video to the TelePresence endpoints. If the TelePresence Bridge detects that the active speaker is from the WebEx side, it switches the TelePresence endpoints to the active speaker video. If the active speaker is from the TelePresence side, the TelePresence Bridge sends the previous active speaker video to the TelePresence endpoint of the current active speaker.

Figure 11-22 Cisco CMR Hybrid Using WebEx Audio with SIP



In the DMZ, Cisco Expressway-E handles the traversal calls between the enterprise and WebEx Cloud, and it allows the signal and media to traverse through the internal and external firewalls. Cisco Expressway-E connects with the WebEx Cloud through the configured DNS Zone and routes calls to WebEx via DNS lookup. Cisco Expressway-E communicates with WebEx Cloud via an encrypted connection using TLS and secured RTP for SIP signal and media. Customers have an option to turn on encryption for the SIP signal and media traffic within the enterprise. TelePresence endpoints outside of the enterprise can register with Unified CM through Expressway-C and Expressway-E, and thus participants on these endpoints can join the CMR Hybrid meetings.

When the WebEx Cloud receives the traversal calls and media sent from the enterprise network, the WebEx audio bridge cascades the audio into the WebEx conference, and WebEx switches to the active speaker video and displays the content sharing on the WebEx meeting clients. Similarly, WebEx Cloud sends the conference mixed audio, the active speaker, and content sharing video from the WebEx side to the enterprise via Cisco Expressway-E and Expressway-C, which routes them to the TelePresence Bridge.

Cisco CMR Hybrid supports H.264 video for active speaker and content sharing. It utilizes Binary Floor Control Protocol (BFCP) for content sharing and G.711 codec for audio. While Cisco WebEx uses H.264 video and G.711 audio codec, TelePresence can still use other video formats or codecs that are supported by the endpoints. The TelePresence Bridge will handle the audio and video interoperability between the TelePresence endpoints and WebEx meeting clients. In addition, there is a flow control on the link between the TelePresence Bridge and WebEx Cloud that regulates the bandwidth available for handling the media. For media from WebEx, the TelePresence Bridge always allocates 4 Mbps to ensure that WebEx sends the best quality of video possible to the TelePresence Bridge. For media from the TelePresence Bridge, the WebEx meeting client has a video floor of 180p for active speaker video at the minimum bit rate of 1.2 Mbps. If the minimum bit rate cannot be maintained due to network conditions (severe packets loss, for example), the WebEx client will stop receiving the active speaker video but still

receives content sharing as well as conference audio and sends its video to other participants. The WebEx client will periodically perform bandwidth retest and automatically reestablish active speaker video when network conditions stabilize. Depending on the capability of the device running the WebEx meeting client and on bandwidth available, the WebEx client supports active speaker video up to HD 720p at 30 frames per second (fps) and content video up to 1080p. During the meeting, WebEx allocates the bandwidth based upon the least capable device among all WebEx clients in the conference (excluding devices running below the video floor), with a maximum bandwidth of 4 Mbps. However, if the least capable device leaves the conference, the bandwidth will be re-allocated based on the next least capable device that runs the WebEx meeting client. The allocated bandwidth determines the resolution and frame rates used to display TelePresence video on WebEx clients. Depending on the TelePresence endpoints deployed, video resolution required, screen layout desired, and deployment options chosen, customers can deploy the TelePresence Bridge using the Cisco TelePresence Server (appliance or virtualized platforms) or Cisco TelePresence MCU, but the pool must consists of bridges of the same type only (either TelePresence Server or TelePresence MCU). For TelePresence Conductor deployment details, see to the section on *Cisco Collaboration Meeting Rooms Premises* in the *Cisco Rich Media Conferencing* chapter of the *Cisco Collaboration System 11.x SRND*, available at

https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cucm/srnd/collab11/collab11/confernc.html

WebEx and TelePresence participants can join the CMR Hybrid meeting from within the enterprise or anywhere from the internet. For WebEx participants, they join the meeting using the WebEx meeting clients with either PSTN or VoIP audio. For TelePresence participants, they join the meeting via the One Button To Push (OBTP) or Auto Connect feature with the supported endpoints or by calling directly into the TelePresence Bridge. Once the participants successfully join the meeting, they can see the live video of each other from the endpoints and meeting clients. For presentation sharing with a WebEx user, either the user can make himself the presenter or the host can assign the presenter privilege to the user before he can start sharing the presentation. There is the WebEx site configuration to control this behavior. For presentation sharing with a TelePresence user, the user can connect the video display cable to his computer or press a button on the endpoint to start sharing his presentation without involving the host.

Note Staring with Cisco TMS 14.6 and TMSPE 1.4, Cisco Collaboration Meeting Rooms Premises can be integrated with Cisco WebEx, allowing participants to join a meeting in the user's personal room from the WebEx meeting client.

Scheduling

Cisco TelePresence Management Suite (TMS) is the key component for scheduling Cisco CMR Hybrid meetings. It provides a control link to the Cisco WebEx meeting scheduler. This link enables Cisco TMS to create new meetings on Cisco WebEx calendar and to obtain Cisco WebEx meeting information that is distributed to meeting participants. The following options are available to schedule CMR Hybrid meetings:

- WebEx Productivity Tools

  WebEx Productivity Tools is a suite of tools that allows users to schedule WebEx sessions quickly and easily. Productivity Tools includes an Outlook plug-in that allows an organizer to schedule WebEx Meetings, TelePresence resources, and CMR Hybrid meetings. Cisco TelePresence Management Suite Extension for Microsoft Exchange (TMSXE) is required for the productivity tool to interface with Cisco TMS for booking the meetings. This option provides a seamless integration for users to schedule CMR Hybrid meetings and to send the invitations to all participants directly inside the email client with a single transaction.

- Smart Scheduler

  Smart Scheduler is a web-based tool that is hosted on Cisco TelePresence Management Suite Provisioning Extension (TMSPE), and it allow users to schedule CMR Hybrid meetings using a browser. This could provide an option for users who would like to schedule meetings on mobile devices.

  Note As long as the Cisco TMSPE option key has been installed, there is no extra license required for using Smart Scheduler.

- WebEx Scheduling Mailbox

  In this option, the network administrator needs to create a special mailbox account in Microsoft Exchange Server. When an organizer schedules a CMR Hybrid meeting, he should include this special mailbox account in the invitees list. Cisco TMSXE monitors this account and requests Cisco TMS to book a CMR Hybrid meeting if it sees this account in the recipients list. This option provides a convenient way, but with limited control of settings, for users to schedule meetings using any email clients that are supported by Exchange, such as Outlook Web Access (OWA).

- Cisco TMS Booking Interface

  With this option, the meeting organizer has to log in to the Cisco TMS portal and schedule the CMR Hybrid meetings from the Booking interface. This interface provides users with control of advanced settings for the meetings, and typically IT or help desk personnel uses this option to schedule meetings.

For Cisco TMS configuration details with these options, refer to the *Cisco Collaboration Meeting Rooms (CMR) Hybrid Configuration Guide*, available at

https://www.cisco.com/en/US/products/ps11338/products_installation_and_configuration_guides_list.html

Scheduling a CMR Hybrid meeting is a two-steps process. First, a request is sent to the WebEx Cloud to schedule the meeting on the WebEx calendar, and the WebEx Cloud responds with the meeting details that are passed to Cisco TMS. Second, Cisco TMS schedules the TelePresence meeting in its calendar. When it is the meeting start time, Cisco TMS pushes the meeting details to the TelePresence Bridge for joining the meeting on WebEx. The meeting details returned from WebEx include the date and time for the meeting, dial-in information, subject, meeting number, URL for joining the meeting, and so forth. Once the meeting has been scheduled, details for the WebEx and TelePresence portions of the meeting are sent to the host, and the host can forward the details to all participants. However, if the productivity tool is used, the meeting details are automatically included in the invitation that the host creates and sends to the meeting participants.

Single Sign On

Cisco CMR Hybrid supports scheduling the WebEx portion of the meeting in Cisco TMS using Single Sign On (SSO). This feature requires the WebEx site to have Cisco TMS provisioned as the delegated partner and to have the Partner Delegated Authentication configured. With SSO enabled in Cisco TMS, only the user's WebEx username is stored in the Cisco TMS user profile without the need of the WebEx password. When the user schedules a CMR Hybrid meeting, WebEx trusts Cisco TMS and requires only the WebEx username stored in Cisco TMS to schedule the meeting in the WebEx calendar. For Cisco TMS configuration details with SSO, refer to the latest version of the *Cisco Collaboration Meeting Rooms (CMR) Hybrid Configuration Guide*, available at

https://www.cisco.com/c/en/us/support/conferencing/telepresence-management-suite-tms/products-installation-and-configuration-guides-list.html

For more information regarding SSO with Cisco WebEx, refer to the white papers and technical notes available at

https://developer.cisco.com/site/webex-developer/develop-test/sso/reference

Security

All communications between the enterprise network and the WebEx Cloud are encrypted (using TLS and secured RTP). Customers also have an option to turn on encryption for the SIP signal and media within the enterprise. A certificate has to be uploaded to the Cisco Expressway-E to ensure that proper handshaking takes place for the TLS connection to be functional. That certificate must be signed by a trusted Root Certificate Authority. For the list of the trusted Root Certificate Authorities, refer to the *Cisco Collaboration Meeting Rooms (CMR) Hybrid Configuration Guide*, available at

https://www.cisco.com/en/US/products/ps11338/products_installation_and_configuration_guides_list.html

A password is required when the TelePresence Bridge calls into WebEx to join the meeting. The password is allocated for each CMR Hybrid meeting scheduled on the WebEx calendar and is embedded in the SIP URI that is returned as part of the meeting details from the WebEx Cloud. This password is encoded into 22 bytes and qualifies for the security standards. At the start of the meeting, the TelePresence Bridge calls into WebEx using this SIP URI, and WebEx validates the password to authorize the call to join the meeting.

Deployment Options

When it is the start time for the CMR Hybrid meeting, Cisco TMS initiates the conference on the TelePresence Bridge through TelePresence Conductor for the TelePresence participants. The TelePresence Bridge makes a SIP call through TelePresence Conductor out to the WebEx Cloud using the SIP URI that was returned as part of the scheduling process and to join the conference on the WebEx side. As a result, the TelePresence Bridge establishes separate audio, active speaker video, and content sharing video streams with the cloud for the meeting. The active speaker video, content sharing video, and conference control always travels over the IP network, but the audio can travel over either the IP network or the PSTN, depending on the deployment options chosen. The various audio options available for CMR Hybrid are:

- WebEx Audio Using SIP, page 11-55, including Cloud Connected Audio
- WebEx Audio Using PSTN, page 11-55
- Teleconferencing Service Provider Audio, page 11-57

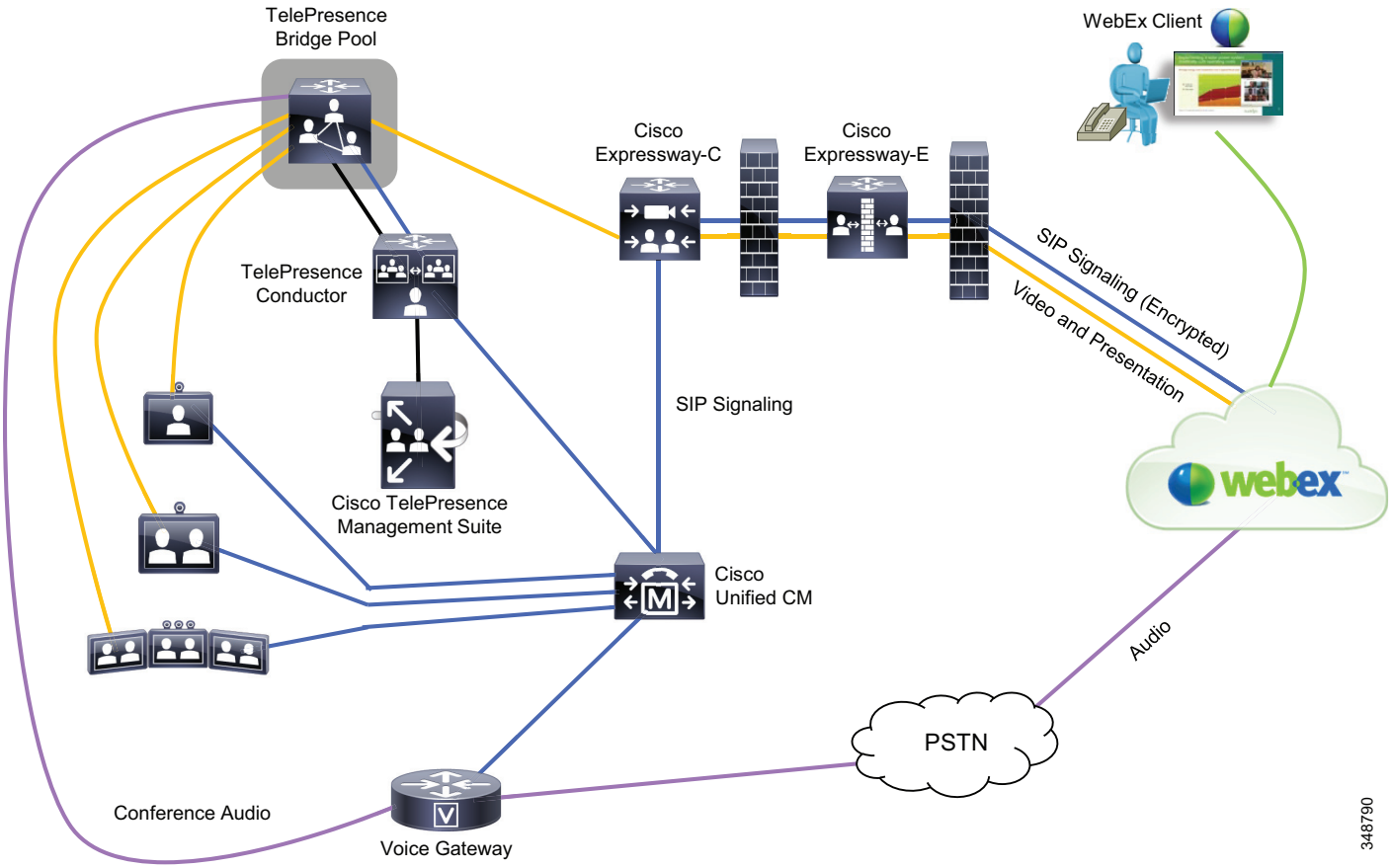
WebEx Audio Using SIP

Figure 11-22 shows the deployment of Cisco CMR Hybrid using WebEx Audio with SIP. In this option, the conference audio is established with the WebEx audio bridge through the SIP connection when the TelePresence Bridge calls out to the WebEx Cloud at the start of the meeting. The audio, active speaker video, content sharing video, and conference control are sent on the IP network from the TelePresence Bridge to the WebEx Cloud through Cisco Expressway-C and Expressway-E. As a result, the audio connection from the TelePresence Bridge cascades into the WebEx audio bridge.

WebEx Audio Using PSTN

For Cisco CMR Hybrid deployment where the in-country rule does not allow toll bypass, WebEx Audio using the PSTN could be an option. Figure 11-23 depicts this deployment. In this option, the active speaker video, content sharing video, and conference control are sent over the IP network, but the audio is established with the WebEx audio bridge through the PSTN. This option requires the deployment of a voice gateway to connect the audio call between the IP network and the PSTN. During the scheduling process, when the meeting is scheduled on the WebEx calendar, WebEx passes the dial-out number and the meeting number to Cisco TMS. At the start of the meeting, the TelePresence Bridge initiates a SIP call to the WebEx Cloud to establish the active speaker video and content sharing video. At the same time, the TelePresence Bridge dials out through the PSTN to establish an audio connection with the WebEx audio bridge. After connecting with the WebEx audio bridge, the TelePresence Bridge sends out the meeting number as a DTMF dial sequence so that WebEx can associate the audio and video call legs. As a result, the audio connection from the TelePresence Bridge cascades into the WebEx audio bridge.

Figure 11-23 Cisco CMR Hybrid Using WebEx Audio with PSTN

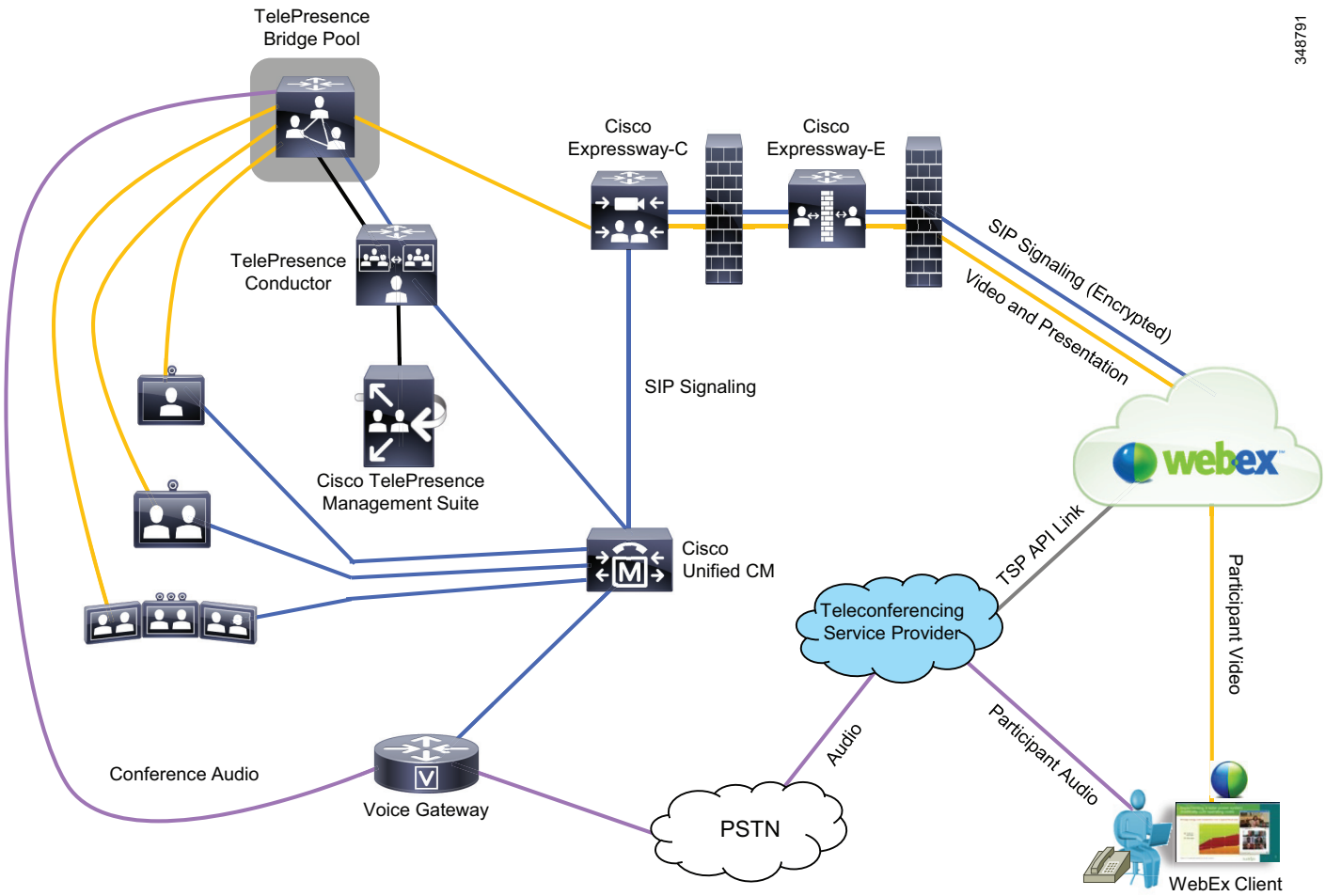


The dial-out number returned from WebEx is in full E.164 number format (for example, +14085551212). The dial plan design in Cisco Unified CM should take into account the handling of E.164 numbers. For dial plan design with Cisco Unified CM, see the chapter on Dial Plan, page 14-1.

Teleconferencing Service Provider Audio

The Teleconferencing Service Provider (TSP) Audio option is for customers who prefer to use the audio bridge hosted by their third-party teleconferencing service provider. The TSP Audio configuration is very similar to WebEx Audio using the PSTN configuration, except that the audio bridge is hosted by the teleconferencing service provider (see Figure 11-24). The TSP link between WebEx and TSP provides the advanced conference control features.

Figure 11-24 Cisco CMR Hybrid Using Teleconferencing Service Provider (TSP) Audio



During the scheduling process, in addition to the dial-out number and meeting number, extra digits for navigating through the IVR prompts on the TSP audio bridge are passed from WebEx to Cisco TMS. At the scheduled meeting start time, the TelePresence Bridge initiates a SIP call to the WebEx Cloud to establish the video connections. At the same time, the TelePresence Bridge dials out to the TSP audio bridge through the PSTN. Then the TelePresence Bridge plays out the meeting number as a DTMF dial sequence, along with additional DTMF digits to navigate through the IVR prompts on the audio bridge to start the meeting. On the WebEx side, WebEx participants start the WebEx session using the meeting client and dial into the TSP audio bridge or have callback from the audio bridge. Thus, the audio streams from TelePresence and WebEx participants are cascaded. From this point onward, information about the loudest speaker, participant list, and so forth in the WebEx side, is passed from the TSP to WebEx through the TSP link and then into the enterprise collaboration network.

The dial-out number returned from WebEx is in full E.164 number format (for example, +14085551212). The dial plan design in Cisco Unified CM should take into account the handling of E.164 numbers. For dial plan design with Cisco Unified CM, see the chapter on Dial Plan, page 14-1.

High Availability

There are two areas that must be considered when designing high availability for CMR Hybrid: the enterprise collaboration network and the WebEx Cloud. The WebEx Cloud is managed by Cisco and already has the redundancy built into the infrastructure. For details, see the section on Cisco WebEx Software as a Service, page 11-26.

In the enterprise collaboration network, utilize the clustering option from Cisco Unified CM and Cisco Expressway to provide redundancy for call control and call routing on the TelePresence endpoints. In case the primary server fails, the backup server can take over the call control and call routing functions. In addition, resiliency of the TelePresence conferencing infrastructure must be considered to handle failure of conference bridges.

For Cisco Unified CM clustering, see the chapter on Call Processing, page 9-1.

For Cisco Expressway clustering, refer to the latest version of the *Cisco Expressway Cluster Creation and Maintenance Deployment Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html

For resiliency of the TelePresence conferencing infrastructure, see the section on Cisco Meeting Server, page 11-7.

Capacity Planning

The WebEx Cloud has the built-in capability to evenly distribute the traffic and dynamically add more capacity if thresholds are exceeded. Capacity planning for Cisco CMR Hybrid involves sizing of the components running within the enterprise. The components include:

- Call Processing Platforms

  Cisco Unified CM must provide enough resources to handle the traffic generated by the TelePresence endpoints. For details, see the section on Capacity Planning for Collaboration Endpoints, page 8-44.

- TelePresence Conferencing

  The Cisco TelePresence Conductor, Cisco TelePresence Server, or Cisco TelePresence MCU must provide enough resources to handle the conference traffic. For details, see the information on capacity planning in the section on *Cisco Collaboration Meeting Rooms Premises* in the *Cisco Rich Media Conferencing* chapter of the *Cisco Collaboration System 11.x SRND*, available at

  https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cucm/srnd/collab11/collab11/confernc.html

- Cisco Expressway

  Cisco Expressway must provide enough resources to handle the traversal call traffic for the deployment. For capacity details, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Network Traffic Planning

Network traffic planning for Cisco CMR Hybrid consists of the following elements:

- WebEx Clients Bandwidth

  The WebEx meeting client uses the Scalable Video Coding (SVC) technology to send and receive video. It uses multi-layer frames to send video and it allows the receiving client to automatically select the best possible resolution to receive video. For more information regarding network traffic planning for WebEx clients, refer to the *Cisco WebEx Network Bandwidth* white paper available at

  https://www.cisco.com/c/en/us/products/collateral/conferencing/webex-meeting-center/white_paper_c11-691351.html

- Bandwidth from Enterprise to WebEx Cloud

  For each call to the WebEx Cloud, a minimum network bandwidth of 1.1 Mbps is required between the enterprise and the WebEx Cloud. For example, if a customer is expecting five simultaneous CMR Hybrid meetings, network bandwidth of 5.5 Mbps is required. At the same time, a maximum bandwidth of 4 Mbps is supported per call.

For optimal SIP audio and video quality between the TelePresence Bridge and the WebEx Cloud, Cisco recommends setting up the video bandwidth of at least 1.3 Mbps in the region associated with each endpoint registering with Cisco Unified CM.

Design Considerations

The following design considerations apply to Cisco CMR Hybrid deployments:

- Upgrade from previous versions of CMR Hybrid that use the Cisco TelePresence MultiPoint Switch infrastructure is not supported, and customers using those previous versions should plan for migration.
- Every user who wants to schedule a CMR Hybrid meeting must have a host account with Cisco TelePresence Session type assigned in the WebEx site.
- Any endpoints that can register with Cisco Unified CM and that are supported by the TelePresence Bridge can be used to join the Cisco CMR Hybrid meeting.
- Only devices managed by the Cisco TelePresence Management Suite (TMS) can use One Button to Push (OBTP) or the Auto Connect feature to join the CMR Hybrid meeting.
- Ensure that the Cisco Unified CM Neighbor Zone in Cisco Expressway-C is configured with Binary Floor Control Protocol (BFCP) enabled.
- Provision Hybrid Audio in the WebEx site to allow the use of SIP audio for the TelePresence Bridge and PSTN audio for WebEx participants.
- Cisco CMR Hybrid does not support Cisco WebEx Meetings Server.
- The TelePresence Bridge becomes the default host if no host is present when it joins the CMR Hybrid meeting, and the host privilege is reassigned to the host when he joins using the WebEx meeting client.
- The TelePresence Bridge will call into the WebEx Cloud at meeting start time even if no TelePresence or WebEx participant has joined yet.
- The organizer's WebEx account and Outlook time zone should match; otherwise, the meeting scheduled in WebEx and in the Cisco TMS calendar will have different start times.
- Enable UDP for media streaming in the firewalls for the optimal video experience.

PART 2

Call Control and Routing

Contents of This Part

This part of the document contains the following chapters:

- Overview of Call Control and Routing
- Bandwidth Management
- Dial Plan
- Emergency Services
- Directory Integration and Identity Management

Overview of Call Control and Routing

Revised: June 15, 2015

Once the network infrastructure has been put in place for your Cisco Unified Communications and Collaboration System, call control and routing applications, components, and services can be layered on top of that infrastructure. There are numerous applications and features that can, and in some cases must, be deployed on the network infrastructure:

- Bandwidth management — Provides mechanisms for ensuring voice and video quality and for preventing oversubscription of network bandwidth by limiting the number of calls that are allowed on the network at a given time. With a combination of packet marking and re-marking, and advance queuing mechanisms such as low latency or priority queuing, voice and video quality is guaranteed. Similarly, call admission control enforces the overall call capacity of the call processing components and available network bandwidth.
- Dial plan — Provides endpoint numbering, dialed digits analysis, and classes of restriction to limit the types of calls that a user can make.
- Emergency services — Provide essential information about the caller's location and emergency situation to the appropriate Public Safety Answering Point (PSAP) so that the caller receives a swift response and the necessary help (for example, police, fire, or ambulance teams).
- Directory and identity management services — Lightweight Directory Access Protocol (LDAP) provides applications with a standard method for accessing and potentially modifying the information stored in a directory. Likewise, identity management and single sign-on ensure that user identification and access are secure. These capabilities enable companies to centralize all user information in a single repository available to multiple applications, resulting in better access to the information and a reduction in maintenance costs through the ease of making adds, moves, and changes.

The chapters in this part of the SRND cover the features, components, and services mentioned above. Each chapter provides an introduction to the component or service, followed by discussions surrounding architecture, high availability, and design considerations. The chapters focus on design-related aspects of the applications and services rather than product-specific support and configuration information, which is covered in the related product documentation.

This part of the SRND includes the following chapters:

- Bandwidth Management, page 13-1

  This chapter examines bandwidth management techniques and the potential for oversubscribing IP links, which causes the voice and video quality of calls to become unacceptable. It also examines the use of call admission control to allow only a certain number of simultaneous calls on the network at a given time to prevent oversubscription. This chapter covers quality of service (QoS) and call admission control types, including location-based call admission control, as well as design and deployment guidelines for successfully deploying QoS and admission control services.

- Dial Plan, page 14-1

  This chapter explores dial plan features and functions that enable the call processing application to route calls to appropriate destinations. The chapter considers various aspects of dial plan services, including dial plan constructs, dial plan numbering options and design considerations, classes of restriction, inbound and outbound calling features, and dial plan and call routing redundancy mechanisms.

- Emergency Services, page 15-1

  This chapter discusses accessing emergency services through Public Safety Answering Points (PSAPs) on the PSTN from within the enterprise IP communications environments, an important aspect of most deployments due to possible critical needs for medical, fire, and other emergency response services. The chapter provides an overview of the various emergency service components both inside and outside the enterprise. It also discusses planning, 911 network service providers, gateway interfaces, and number-to-location mapping.

- Directory Integration and Identity Management, page 16-1

  This chapter covers aspects of Unified Communications and Collaboration integration with the LDAP directories, including the Cisco Unified Communications Manager directory architecture itself, as well as design considerations for LDAP synchronization and authentication. Directory access from Unified Communications and Collaboration endpoints as well as security considerations such as single sign-on are also explored.

Architecture

Call routing components and services such as call processing agents and IP and PSTN gateways rely on the underlying network infrastructure for network connectivity and access. By connecting to the underlying network infrastructure, call routing components and features are able to leverage end-to-end network connectivity and quality of service to access both the enterprise and public telephone networks. In turn, call routing applications and services provide basic Unified Communications and Collaboration functions such as call control, dial plan, call admission control, and gateway services to other applications and services in the deployment. For example, a Unified CM cluster connects to the IP network through a switch in order to communicate with other devices and applications within the network as well as to access other devices and services in other locations. At the same time, the Unified CM cluster provides services such as phone registration and media resource provisioning and allocation to call control components and services such as IP phones.

Further, just as call routing components rely on the network infrastructure for network connectivity, call routing components and services are also often dependent upon each other for full functionality. For example, while Unified CM provides registration and call routing services to various IP endpoints within the network, it is completely dependent upon gateways and gateway services to route calls beyond the enterprise.

High Availability

As with the network infrastructure, critical Unified Communications and Collaboration call routing services should be made highly available to ensure that required features and functionality remain available if failures occur in the network or with individual call routing components. It is important to understand the various types of failures that can occur and the design considerations around those failures. In some cases, the failure of a single server or component (for example, a subscriber node in a Unified CM cluster) might have little or no impact due to the redundant nature of the Unified CM clustering mechanism. However, in other cases a single failure can impact multiple components or services. For example, the failure of a PSTN or IP gateway could result in loss of access to the public telephone network, and even though a call processing agent such as Unified CM is still available and able to provide most features and services, it cannot route calls to the PSTN because there is no path available if the gateway fails. To avoid these types of situations, you should deploy multiple PSTN gateways to provide redundant gateway services, and you should configure the call processing agent to handle call routing to both gateways as needed.

For features and services such as dial plan and bandwidth management, high availability considerations include temporary loss of functionality due to network connectivity or call processing agent application server failures, which result in the inability of the call agent to route calls and therefore the inability for callers to make calls. Oversubscription of the network could also occur if QoS and other call admission control services are not available to the endpoints initiating calls. For example, if the call admission control agent fails or loses connectivity to the network, the call may still go through but without the call admission control service being aware of the call, thus potentially resulting in poor quality. To avoid these types of scenarios, provide call admission control resiliency by deploying multiple call admission control agents.

High availability considerations are also a concern for components and services such as video endpoints and remote site survivability. For deployments with network-attached remote sites where devices are leveraging call processing services from an agent in a central site, remote site survivability using SRST, for example, can ensure that local phones within the remote site will still receive call processing services in the event of a connectivity failure to the central site. Likewise, to ensure that video endpoints are highly available, you can deploy more than one multipoint control unit (MCU) in case one fails.

Capacity Planning

The network infrastructures must be designed and deployed with consideration for the capacity and scalability of the individual components and the overall system. Similarly, deployments of call routing components and services must also be designed with attention to capacity and scalability considerations. When deploying various call routing applications and services, not only is it important to consider the scalability of the applications and services themselves, but you must also consider the scalability of the underlying network infrastructure. Certainly the network infrastructure must have available bandwidth and be capable of handling the additional traffic load that the call routing components will create. Similarly, the call routing infrastructure and its components must be capable of handling all the required device configurations and registrations as well as the call load or busy hour call attempts (BHCA),

For example, with call processing agents such as Unified CM, it is critical to assess the size of the deployment in terms of number of users, endpoints, and calls per user per hour, and to deploy sufficient resources to handle the required load. If a call processing agent is undersized and does not have sufficient resources, features and services will begin to fail as the load increases. Two of the chief considerations when attempting to size a call processing deployment are the call processing type and the call processing hardware. Both of these are critical for sizing the system appropriately given the number of users, locations, devices, and so on. As an example, Cisco Unified Communications Manager has a much

higher capacity than Cisco Unified Communications Manager Express and should therefore be used for larger deployments. In addition, the server platform selected to run the call processing agent will, in many cases, determine the maximum load.

Capacity planning for remote site survivability is much the same in that it relies on backup call processing hardware. Selecting the appropriate Cisco IOS platform to provide backup or survivable call processing services typically begins with determining the number of devices and users that must be supported at that site in the event that connectivity to the central site is disrupted. Equally critical in this sizing exercise are the local PSTN gateway services. In the event of a central site connection failure, will the local PSTN gateway have sufficient circuits to be able to route all calls without blocking during the busiest hour? If the answer is no, adding more gateways or trunks will be necessary to size the remote site appropriately for backup call processing.

PSTN and IP gateways must also be sized appropriately for a deployment, so that sufficient capacity is available to handle all calls in the busiest hour. In some cases, you might have to deploy multiple PSTN or IP gateways to provide enough resources.

When configuring and sizing QoS and call admission control services, ensure that sufficient bandwidth is available over network connections and in priority queues to support the required number of calls. If sufficient bandwidth is not available, additional network capacity, gateways, and IP or telephony trunks might be required.

Sizing dial plan services is also important. However, in most cases dial plan capacity in terms of the number of endpoints or phone numbers, route patterns, or other dial plan constructs, is completely dependent upon the type of call processing agent and platform used.

For components and services such as video telephony, appropriate sizing is just as critical. Capacity planning considerations for video telephony focus mainly on network bandwidth, available video ports, and MCU sessions. In most cases additional capacity can be added by increasing the number of application servers and MCUs or by upgrading server or MCU hardware with higher-scale models, assuming the underlying network infrastructure is capable of handling the additional load.

For a complete discussion of system sizing, capacity planning, and deployment considerations related to sizing, refer to the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Bandwidth Management

Revised: March 1, 2018

Bandwidth management is about ensuring the best possible user experience end-to-end for all voice and video capable endpoints, clients, and applications in the Collaboration solution. This chapter provides a holistic approach to bandwidth management that incorporates an end-to-end Quality of Service (QoS) architecture, call admission control, and video rate adaptation and resiliency mechanisms to ensure the best possible user experience for deploying pervasive video over managed and unmanaged networks.

This chapter starts with a discussion of collaboration media and the differences between audio and video, and the impact that this has on the network. Next an end-to-end QoS architecture for collaboration is discussed, with techniques for how to identify and classify collaboration media and signaling for both trusted and untrusted endpoints, clients, and applications. WAN queuing and scheduling strategies are also covered, as well as bandwidth provisioning and admission control.

Note The chapter on Network Infrastructure, page 3-1, lays the foundation for QoS in the LAN and WAN. It is important to read that chapter and fully understand the concepts discussed therein. This chapter assumes an understanding of those concepts.

What's New in This Chapter

Table 13-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

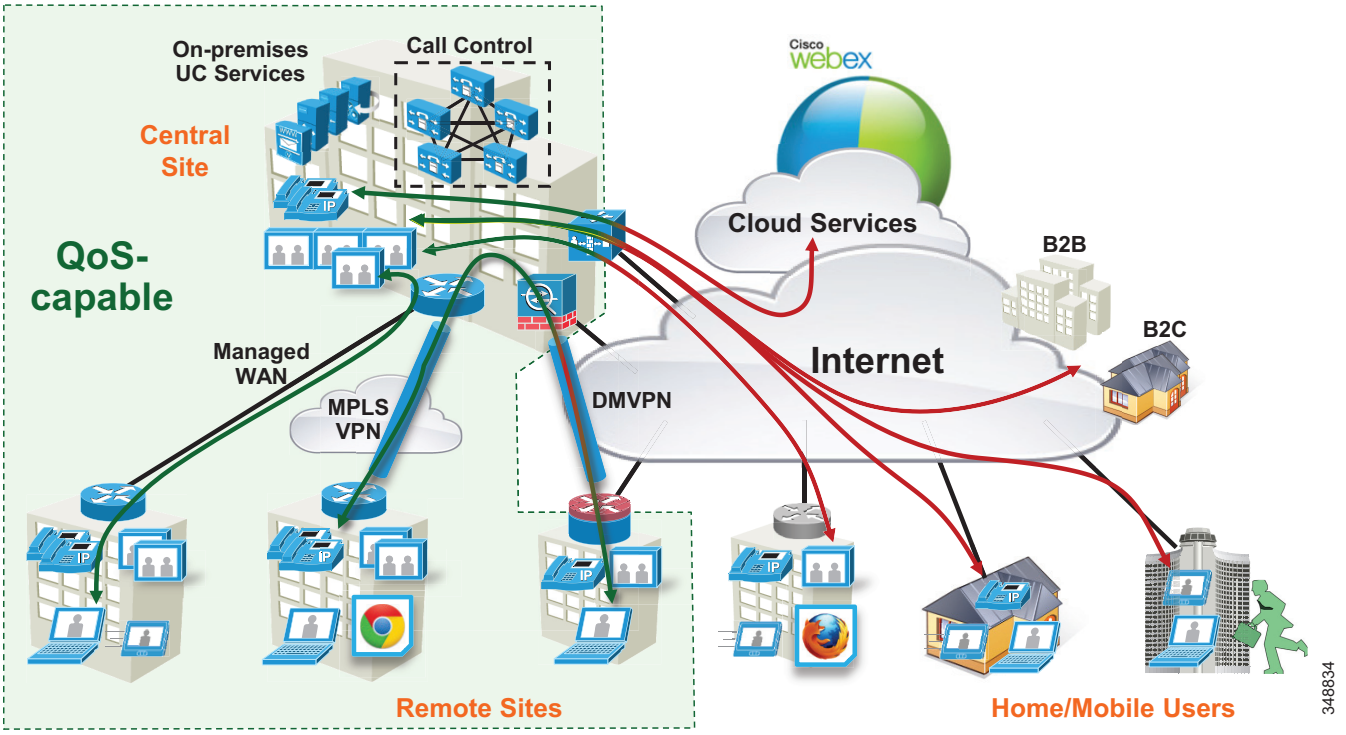
Table 13-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
QoS policy matching and DSCP re-marking	Trusted Endpoints, page 13-93	March 1, 2018
Other minor corrections and updates	Various sections of this chapter	March 1, 2018

Introduction

The collaboration landscape is constantly evolving, and two areas that have changed dramatically are the applications and the network. When Unified Communications was first introduced, it consisted primarily of fixed hardware endpoints such as IP phones and room system endpoints connected to a completely managed network where the administrators were able to implement Quality of Service (QoS) everywhere throughout the network where media traversed. Over time, usage of the Internet and cloud-based services such as WebEx have been added, which means that some of the collaboration infrastructure is now located outside of the managed network and in the cloud. The office connectivity options have also evolved, and companies are interconnecting remote sites and mobile users over the Internet either directly connected over Cisco Expressway, for example, or over technologies such as Dynamic Multipoint VPN (DMVPN). Figure 13-1 illustrates the convergence of a traditional on-premises Collaboration solution in a managed (capable of QoS) network with cloud services and sites located over an unmanaged (not capable of QoS) network such as the Internet. On-premises remote sites are connected over this managed MPLS network where administrators can prioritize collaboration media and signaling with QoS, while other remote sites and branches connect into the enterprise over the Internet, where collaboration media and signaling cannot be prioritized or can be prioritized only outbound from the site. Many different types of mobile and teleworkers also connect over the Internet into the on-premises solution. So the incorporation of the Internet as a source for connecting the enterprise with remote sites, home and mobile users, as well as other businesses and consumers, is becoming pervasive and has an important impact on bandwidth management and user experience.

Figure 13-1 Managed versus Unmanaged Network

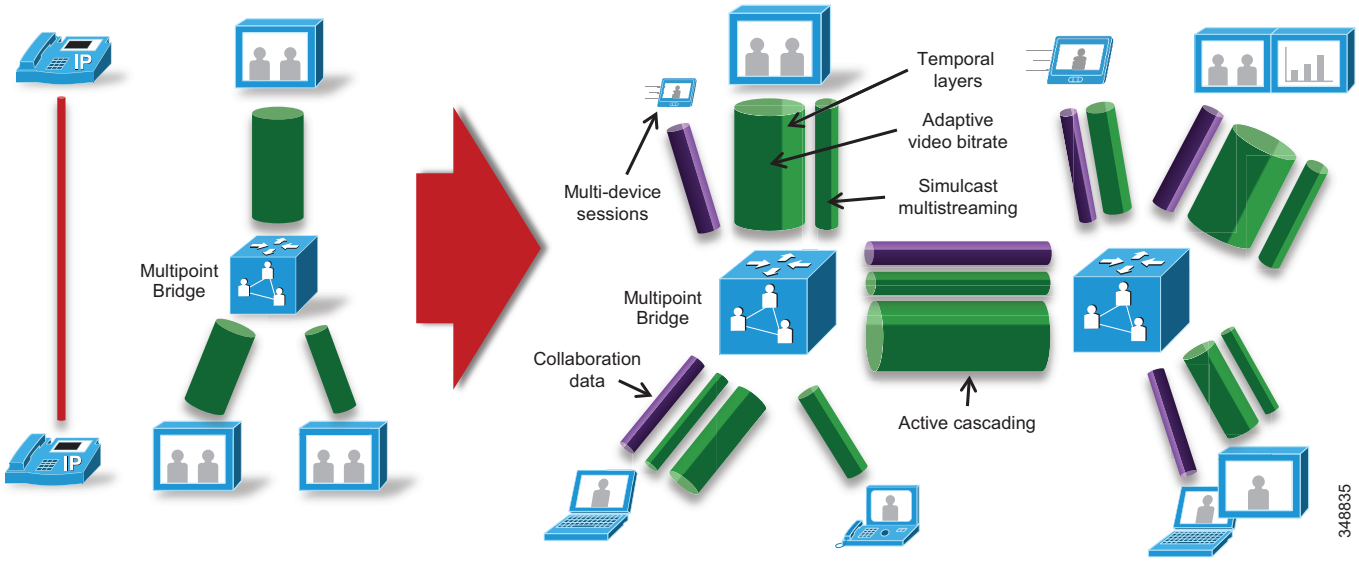


New technologies and trends also mean an evolution of endpoints and user experiences and a plethora of collaboration devices and options. The enterprise is moving from housing single-purpose, single-media communications devices to multi-purpose, multi-media options. This is evident in trends

such as Bring Your Own Device (BYOD), where users are bringing to the enterprise their compact and powerful mobile devices and incorporating collaboration technologies such as instant messaging, video collaboration and conferencing, and desktop sharing, to name a few, into their work processes, making them more collaborative and efficient.

Collaboration media has also greatly evolved from fixed single-stream, fixed bit rate audio and video streams connected point-to-point or via a multi-point bridge to multi-layer, multi-stream, adaptive bit rate video sessions cascaded across multi-point bridges interconnecting a variety of devices. Figure 13-2 illustrates this evolution.

Figure 13-2 The Evolution of Collaboration Media Streams



Other technologies and trends that are currently and actively being adopted in the collaboration solution include:

- Mobility, Bring Your Own Device (BYOD), and ubiquitous video
- Web-based collaboration and WebRTC
- Standard versus immersive video
- Cloud, on-premises, and hybrid conferencing
- Wide-area networks: owned versus over-the-top
- Inter-company collaboration: business-to-business and business-to-consumer
- Multi-device, multi-stream sessions: voice, video, data sharing, and instant messaging

This evolution of managed versus unmanaged networks, new endpoints, and user experiences as well as new technologies and trends have brought with them challenges such as:

- How to manage the bandwidth and ensure a high-quality user experience over managed and unmanaged networks
- How to deploy video pervasively across the enterprise and optimize bandwidth utilization of the available network resources

This chapter presents a strategy of leveraging smart media techniques in Cisco Video endpoints, building an end-to-end QoS architecture, and using the latest design and deployment recommendations and best practices for managing bandwidth to achieve the best user experience possible based on the network resources available and the types of networks collaboration media are now forced to traverse.

Collaboration Media

This section covers the characteristics of audio and video streams in real-time media, as well as the smart media techniques that Cisco Video endpoints employ to ensure high fidelity video in the face of packet loss, delay, and jitter.

Fundamentals of Digital Video

Video is a major component of the enterprise traffic mix. Both streaming and pre-positioned video have implications on the network that can substantially affect overall performance. Understanding the structure of video datagrams and the requirements they place on the network can assist network administrators with implementing a media-ready network.

Different Types of Video

There are several broad attributes that can be used to describe video. For example, video can be categorized as real-time or prerecorded, streaming or pre-positioned, and high resolution or low resolution. The network load is dependent on the type of video being sent. Prerecorded, pre-positioned, low resolution video is little more than a file transfer, while real-time streaming video demands a high-performance network. Many generic video applications fall somewhere in between. This allows non-real-time streaming video applications to work acceptably over the public Internet. Tuning the network and media encoders is an important aspect of deploying video on an IP network.

H.264 Coding and Decoding Implications

Video codecs have been evolving over the last 15 years. Today's codecs take advantage of the increased processing power to better optimize the stream size. The general procedure has not changed much since the original MPEG1 standard was released. Pictures consist of a matrix of pixels that are grouped into blocks. Blocks combine into macro blocks. A row of macro blocks is a slice. Slices form pictures, which are combined into groups of pictures (GOPs).

Each pixel has a red, green, and blue component. The encoding process starts by color sampling the RGB into a luma and two-color components, commonly referred to as YCrCb. Small amounts of color information can be ignored during encoding and then replaced later by interpolation. Once in YCrCb form, each component is passed through a transform. The transform is reversible and does not compress the data. Instead, the data is represented differently to allow more efficient quantization and compression. Quantization is then used to round out small details in the data. This rounding is used to set the quality. Reduced quality allows better compression. Following quantization, lossless compression is applied by replacing common bit sequences with binary codes. Each macro block in the picture goes through this process, resulting in an elementary stream of bits. This stream is sliced into 188-byte packets known as a Packetized Elementary Stream (PES). This stream is then loaded into IP packets. Because IP packets have a 1,500 byte MTU and PES packets are fixed at 188 bytes, only 7 PES can fit into an IP packet. The resulting IP packet will be 1,316 bytes, not including headers. As a result, IP fragmentation is not a concern. An entire frame of high definition video may require 100 IP packets to carry all of the elementary stream packets, although 45 to 65 packets are more common. Quantization

and picture complexity are the primary factors in determining the number of packets required for transmission. Forward error correction can be used to estimate some lost information. However, in many cases multiple IP packets are dropped in sequence. This makes the frame almost impossible to decompress. The packets that were successfully sent represent wasted bandwidth. RTCP can be used to request a new frame. Without a valid initial frame, subsequent frames will not decode properly.

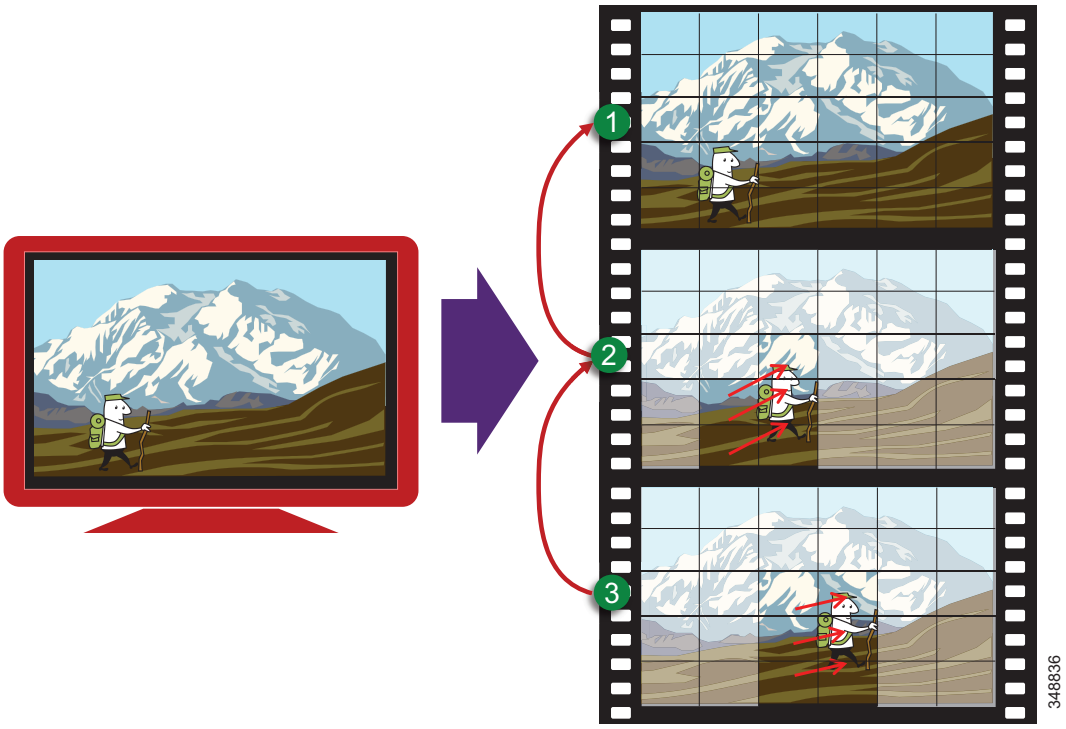
Frame Types

The current generation of video coding is known by three names; H.264, MPEG4 part 10, and Advanced Video Coding (AVC). As with earlier codecs, H.264 employs spatial and temporal compression. Spatial compression is used on a single frame of video as described previously. These types of frames are known as I-frames. An I-frame is the first picture in a GOP. Temporal compression takes advantage of the fact that little information changes between subsequent frames. Changes are a result of motion, although changes in zoom or camera movement can result in almost every pixel changing. Vectors are used to describe this motion and are applied to a block. A global vector is used if the encoder determines all pixels moved together, as is the case with camera panning. In addition, a difference signal is used to fine-tune any error that results. H.264 allows variable block sizes and is able to code motion as fine as ¼ pixel. The decoder uses this information to determine how the current frame should look based on the previous frame. Packets that contain the motion vectors and error signals are known as P-frames. Lost P-frames usually results in artifacts that are folded into subsequent frames. If an artifact persists over time, then the likely cause is a lost P-frame.

Figure 13-3 illustrates how this works in a basic manner:

1. An I-frame (Intra-coded picture) is the entire picture encoded as a static image and sent as a group of packets. This frame does not reference any other frame, and the decoder requires only this frame to build the entire image. In this case the image is of a little hiker hiking through the mountains.
2. Next a P-frame (Predicted picture) is sent, which is a frame based on a previously encoded frame (in this case the I-frame), and only the differences from that I-frame are encoded. The decoder takes these differences and applies them to the I-frame that it had. In this case it shows the little hiker moving up the hill. Because only the little hiker and his movement have changed from the last I-frame, this P-frame is much smaller and represents fewer packets and thus less bandwidth to be transmitted.
3. The next P-frame is sent and is a prediction from the last P-frame sent. As in the P-frame from step 2, this P-frame shows the difference between the last movement of the hiker up the hill and this new movement of the hiker. This progression continues until there is a larger amount of change from the previous image, thus requiring a new I-frame.

Figure 13-3 Encoding Basics



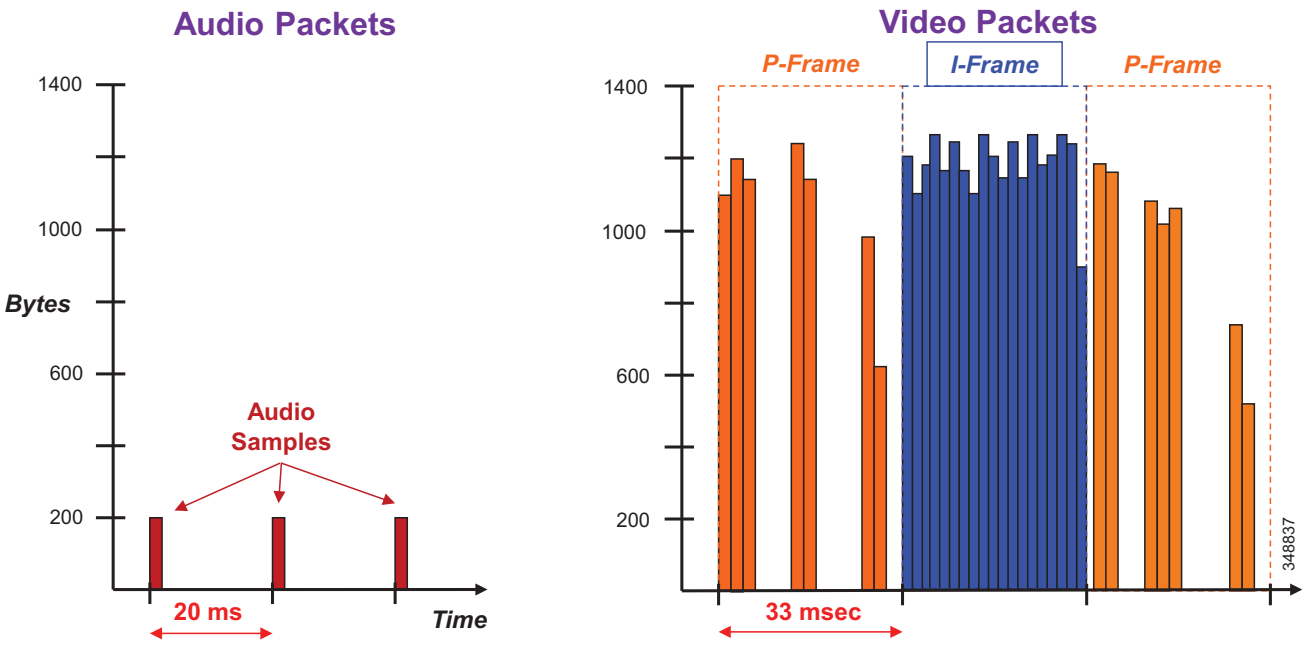
H.264 also implements B-frames. This type of frame fills in information between P-frames. This means that the B-frame must be held until the next P-frame arrives, before the B-frame information can be used. B-frames are not used in all modes of H.264. The encoder decides what type of frame is best suited. There are typically more P-frames than I-frames. Lab analysis has shown TelePresence I-frames to generally be 64 Kbytes wide (50 packets @ 1,316 bytes), while P-frames average 8 Kbytes wide (9 packets at 900 bytes). So I-frames are larger and create the spikes in bit rate in comparison to P-frames.

Audio versus Video

Voice and video are often thought of as close cousins. Although they are both real-time protocol (RTP) applications, the similarities stop there. Voice is generally considered well behaved because each packet is a fixed size and fixed rate. Video frames are spread over multiple packets that travel as a group. Because one lost packet can ruin a P-frame, and one bad P-frame can cause a persistent artifact, video generally has a tighter loss requirement than audio. Video is asymmetrical. Voice can also be asymmetrical but typically is not. Even on mute, an IP phone will send and receive the same size flow.

Video increases the average real-time packet size and has the capacity to quickly alter the traffic profile of networks. Without planning, this could be detrimental to network performance. Figure 13-4 shows the difference between a series of audio packets and a series of video packets sent over a specific time interval.

Figure 13-4 Audio versus Video



As can be seen from Figure 13-4, the audio packets are the same size, sent at exactly the same time intervals, and they represent a very smooth stream. Video, on the other hand, sends a larger group of packets over fixed intervals and can vary greatly from frame to frame. Figure 13-4 shows the difference in the number of packets and packet sizes for an I-frame as opposed to P-frames. This translates to a stream of media that is very bursty in nature when compared to audio. Figure 13-5 illustrates the bandwidth profile over time of an HD video stream. Note the large bursts when I-frames are sent.

Figure 13-5 Bandwidth Usage: High-Definition Video Call

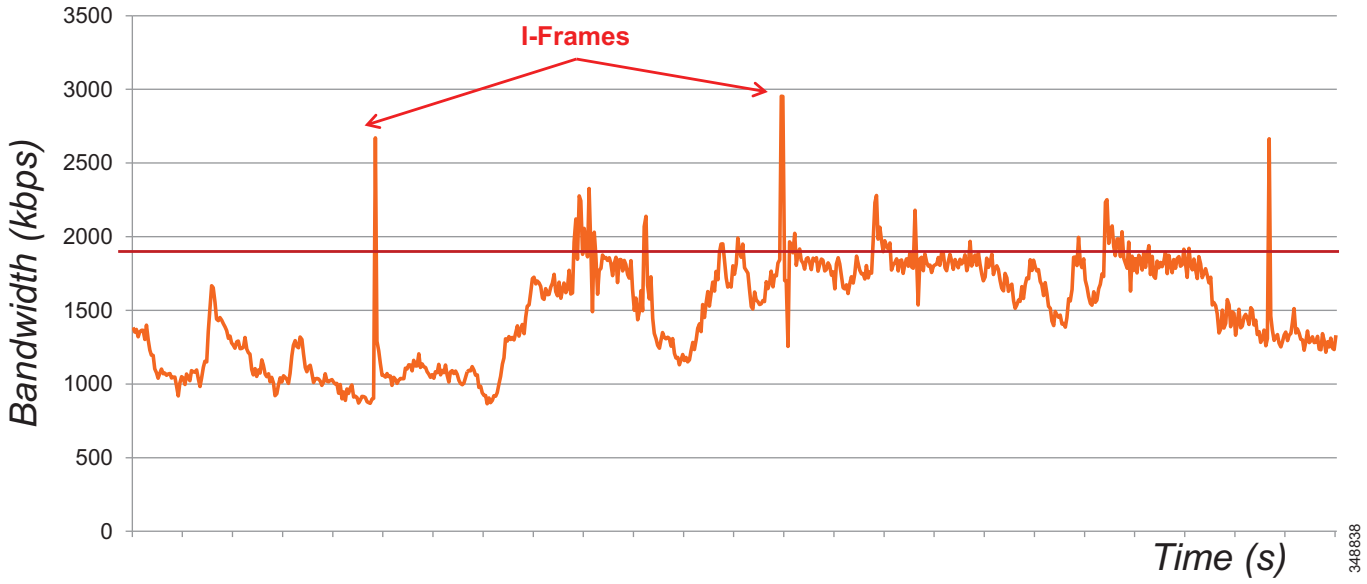
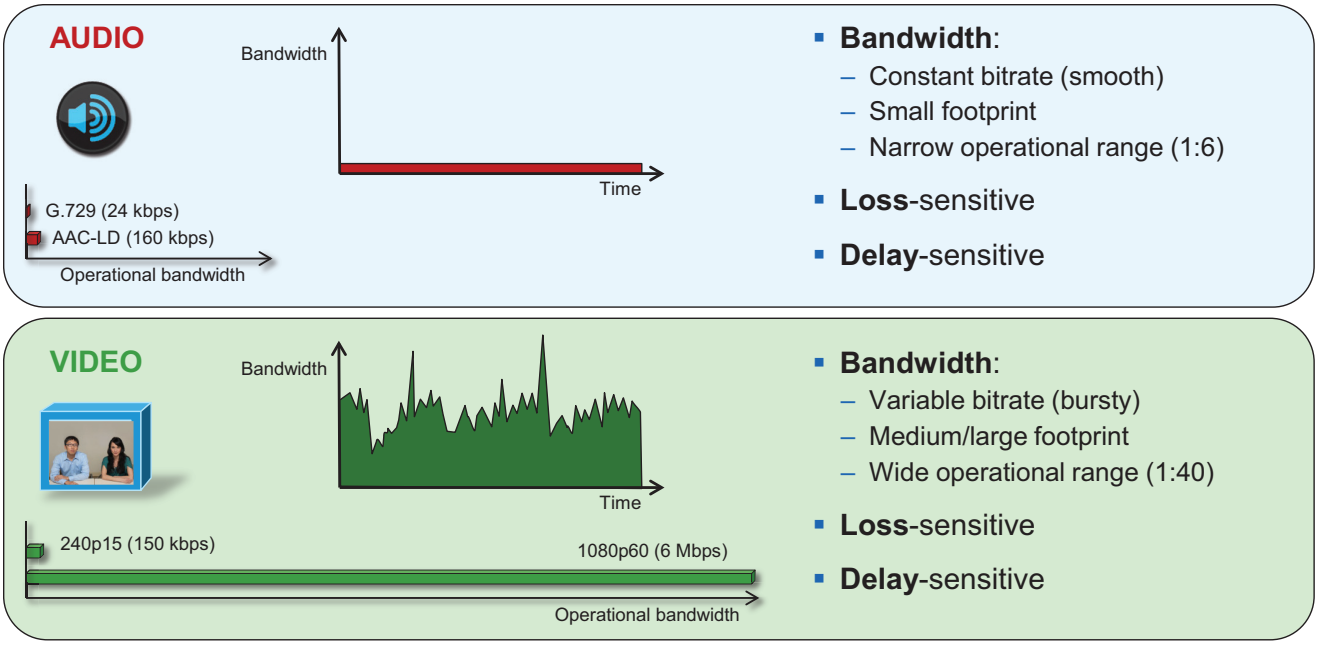


Figure 13-5 shows an HD video call, 720p30 @ 1,920 kbps (1,792 kbps video + 128 kbps audio). The graph shows the video bandwidth (including L3 overhead), and the red line indicates average bit rate.

While audio and video are both transported over UDP and sensitive to loss and delay, they are quite different in their network requirements and profile. Audio is a constant bit rate and has a smaller footprint compared to video, as well as a narrower operational range of 1:6 ratio when comparing the lowest bit-rate audio codec to one of the highest bit-rate codecs. Video, on the other hand, has a variable bit rate (is bursty) and has a medium to large footprint when compared to audio, as well as a wide operational range of 1:40 (250p at 15 fps vs 1080p at 60 fps). Figure 13-6 illustrates some of these differences.

Figure 13-6 Video Traffic Requirements and Profiles



The important point to keep in mind is that audio and video, while similar in transport and sensitivity to loss and delay, are quite different with regard to managing their bandwidth requirements in the network. It should also be noted that, while video is pertinent to a full collaboration experience, audio is critical. If, for example, video is lost during a video call due to a network outage or some other network related event, communication can continue provided that audio is not lost during this outage. This is a critical concept when thinking through the network requirements of a collaboration design such as QoS classification and marking.

Resolution

The sending station determines the video resolution and, consequently, the load on the network. This is irrespective of the size of the monitor used to display the video. Observing the video is not a reliable method to estimate load. Common high definition formats are 720i, 1080i, 1080p, and so forth. In addition to high resolution, there is also a proliferation of lower quality video that is often tunneled in HTTP (or in some cases HTTPS) and SSL (see Table 13-2). Typical resolutions include CIF (352x288) and 4CIF (704x576). These numbers were chosen as integers of the 16x16 macro blocks that are used by the DCT (22x18) and (44x36) macro blocks respectively.

Table 13-2 Format, Resolution, and Bandwidth

Format	Resolution	Typical Bandwidth
QCIF (1/4 CIF)	176x144	260 kbps
CIF	352x288	512 kbps
4CIF	704x576	1 Mbps
SD NTSC	720x480	Analog, 4.2 MHz
720 HD	1280x720	1 to 8 Mbps
1080 HD	1080x1920	5 to 8 Mbps H.264 12+ Mbps MPEG-2

Network Load

The impact of resolution on the network load is generally a squared factor; an image that is twice as big will require four times the bandwidth. In addition, the color sampling, quantization, and frame rate also impact the amount of network traffic. Standard rates are 30 frames per second (fps), but this is an arbitrary value chosen based on the frequency of AC power. In Europe, analog video is traditionally 25 fps. Cineplex movies are shot at 24 fps. As the frame rate decreases, the network load also decreases and the motion becomes less life-like. Video above 24 fps does not noticeably improve motion.

The sophistication of the encoder also has a large impact on video load. H.264 encoders have great flexibility in determining how best to encode video, and with this comes complexity in determining the best method. For example, MPEG4.10 allows the encoder to select the most appropriate block size depending on the surrounding pixels. Because efficient encoding is more difficult than decoding, and because the sender determines the load on the network, low-cost encoders usually require more bandwidth than high-end encoders. H.264 coding of real-time CIF video will drive all but the most powerful laptops well into 90% CPU usage without dedicated media processors.

Table 13-3 through Table 13-5 show average bandwidth utilization ranges based on endpoint and resolution. These tables are provide only as an example of the bandwidth ranges based on resolution of common TelePresence and desktop video endpoints. Refer to the current product documentation for the latest numbers relevant to the endpoints in question.

Table 13-3 Cisco TelePresence Endpoints – Example Bandwidth Usage[1]

	MX200		SX20		EX90		TX9000	
Resolution	Lowest	Highest	Lowest	Highest	Lowest	Highest	Lowest	Highest
720p30 (1280x720)	736 kbps	1.2 Mbps	812 kbps	1.2 Mbps	812 kbps	1.2 Mbps	3.1 Mbps	6.4 Mbps
1080p30 (1920x1080)	2.6 Mbps	5.7 Mbps	2.6 Mbps	6.2 Mbps	2.5 Mbps	6.1 Mbps	8.8 Mbps	11.9 Mbps
720p60 (60 fps)	N/A	2.3 Mbps	N/A	2.3 Mbps	N/A	2.4 Mbps	N/A	N/A

1. For more information on TelePresence endpoints, refer to the bandwidth usage white paper available at https://www.cisco.com/c/dam/en/us/products/collateral/collaboration-endpoints/tested_bandwidth_whitepaperx.pdf.

Table 13-4 Cisco DX Series – Example Bandwidth Usage[1]

Resolution	DX Series Video Bandwidth
240p30 (432x240)	150 to 299 kbps
360p30 (640x360)	300 to 599 kbps
480p30 (848x480)	600 to 799 kbps
576p30 (1024x576)	800 kbps to 1.29 Mbps
720p30 (1280x720)	1.3 to 1.99 Mbps
1080p30 (1920x1080)	2 to 4 Mbps

1. For more information on the DX Series, refer to the latest version of the *Cisco DX Series Administration Guide*, available at https://www.cisco.com/c/en/us/support/collaboration-endpoints/desktop-collaboration-experience-dx600-series/products-maintenance-guides-list.html.

Table 13-5 Cisco Jabber – Example Bandwidth Usage[1]

Resolution	Jabber Video Bandwidth (with G.711 audio)
w144p30 (256x144)	156 kbps
w288p30 (512x288)	320 kbps
w448p30 (768x448)	570 kbps
w576p30 (1024x576)	890 kbps
720p30 (1280x720)	1.3 Mbps

1. For more information on Jabber, refer to the latest version of the Cisco Jabber Deployment and Installation Guide, available at https://www.cisco.com/c/en/us/support/unified-communications/jabber-windows/products-installation-guides-list.html.

Multicast

Broadcast video lends itself well to taking advantage of the bandwidth savings offered by multicast. This has been in place in many networks for years. Recent improvements to multicast simplify the deployment on the network. Multicast will play a role going forward; however, multicast is not used in all situations. Some applications such as multipoint TelePresence use a dedicated MCU to replicate video. The MCU can make decisions concerning which participants are viewing each sender. The MCU can also quench senders that are not being viewed.

Transports

MPEG4 uses the same transport as MPEG2. A PES consists of 188-byte datagrams that are loaded into IP. The video packets can be loaded into RTP/UDP/IP or HTTP(S)/TCP/IP.

Video over UDP is found with dedicated real-time applications such as multimedia conferencing and TelePresence. In this case, an RTCP channel can be set up from the receiver toward the sender. This is used to manage the video session. RTCP can be used to request I-frames or report capabilities to the sender. UDP and RTP each provide a method to multiplex channels. Audio and video typically use different UDP ports but also have unique RTP payload types. Deep packet inspection (DPI) can be used on the network to identify the type of video and audio that is present. Note that H.264 also provides a mechanism to multiplex layers of the video.

Buffering

Jitter and delay are present in all IP networks. Jitter is the variation in delay. Delay is generally caused by interface queuing. Video decoders can employ a play-out buffer to smooth out jitter found in the network. There are limitations to the depth of this buffer. If it is too small, then drops will result. If it is too deep, then the video will be delayed, which could be a problem in real-time applications such as TelePresence. Another limitation is handling dropped packets that often accompany deep play-out buffers. If RTCP is used to request a new I-frame, then more frames will be skipped at the time of re-sync. The result is that dropped packets have a slightly greater impact in video degradation than they would have if the missing packet had been discovered earlier. Most codecs employ a dynamic play-out buffer.

Summary

Video can dramatically impact the performance of the network if planning does not properly account for this additional load. This chapter attempts to assist administrators in managing real-time video in enterprise networks.

"Smart" Media Techniques (Media Resilience and Rate Adaptation)

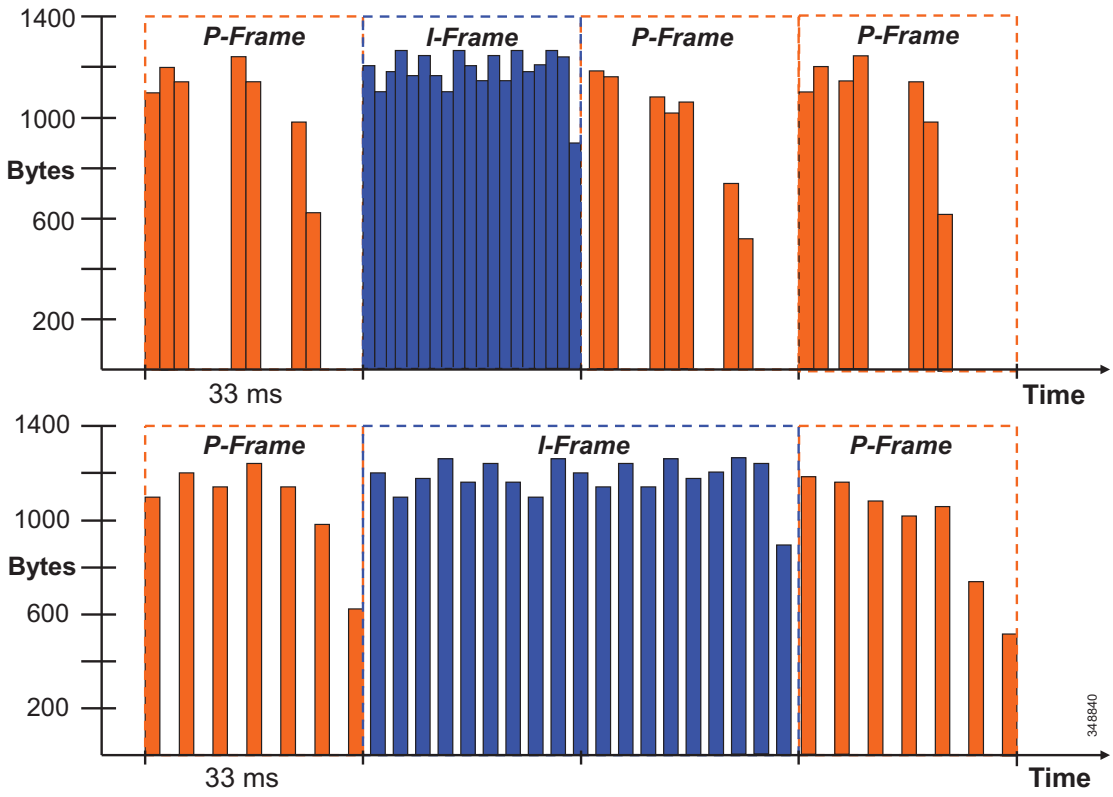
Cisco enterprise video endpoints have evolved greatly over the last few years. Every Cisco video endpoint employs a number of media resiliency techniques to avoid network congestion, recover from packet loss, and optimize network resources. This section covers the following smart media techniques employed by Cisco video endpoints:

- Encoder Pacing, page 13-11
- Gradual Decoder Refresh (GDR), page 13-12
- Long Term Reference Frame (LTRF), page 13-13
- Forward Error Correction (FEC), page 13-14
- Rate Adaptation, page 13-15

Encoder Pacing

The number of packet can increase dependent on the frame type (I or P) as well as the number of packets required, which means that bursts of packets can show up at the beginning, middle, or end of a 33 ms time interval. This creates spikes in bandwidth as the packets are put onto the wire. Encoder pacing is a simple technique used to spread the packets as evenly as possible across the 33 ms interval in order to smooth out the peaks of the bursts of bandwidth. Figure 13-7 illustrates this technique.

Figure 13-7 Encoder Pacing



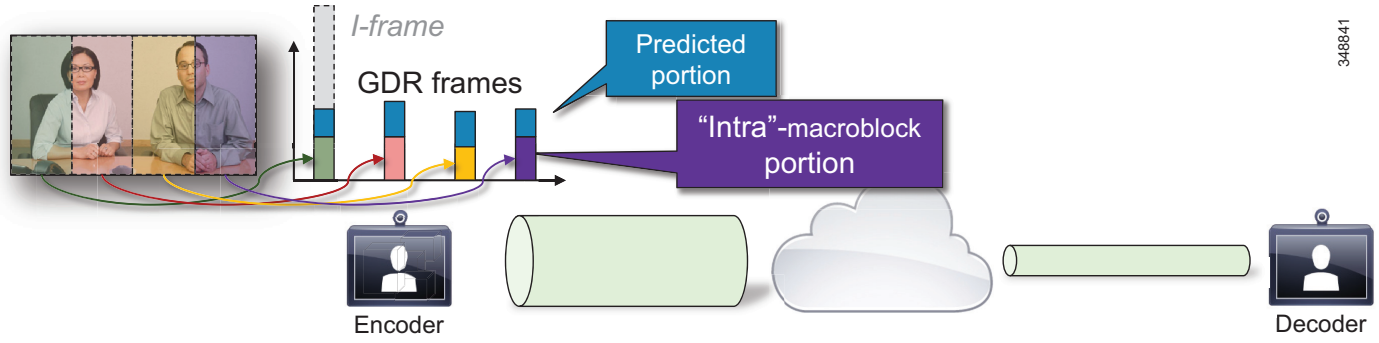
The top image in Figure 13-7 shows packets being placed on the wire without encoder pacing, and the bottom image is with encoder pacing. As each frame is packetized onto the wire in a 33 ms interval, an endpoint packet scheduler disperses packets as evenly as possible across that single interval. Large I-frames might have to be "spread" over two or three frame intervals, and the encoder might then skip one or two frames to stay within a bit rate budget. This smooths out the peaks in bandwidth utilization over the same time frame.

Gradual Decoder Refresh (GDR)

GDR provides a starting point or refresh of the encoded bit stream. GDR is a method of gradually refreshing the picture over a number of frames, giving a smoother and less bursty bit stream.

A new I-frame causes a traffic burst, which in turn can generate congestion, particularly in switched conferences. If one I-frame packet gets dropped, the whole frame needs to be retransmitted. As illustrated in Figure 13-8, Gradual Decoder Refresh spreads "intra"-encoded picture data over N frames. The GDR frames contain a portion of "intra" macroblocks and a portion of predicted macroblocks. Once all GDR frames have been received, the decoder can fully refresh the picture.

Figure 13-8 Gradual Decoder Refresh (GDR)



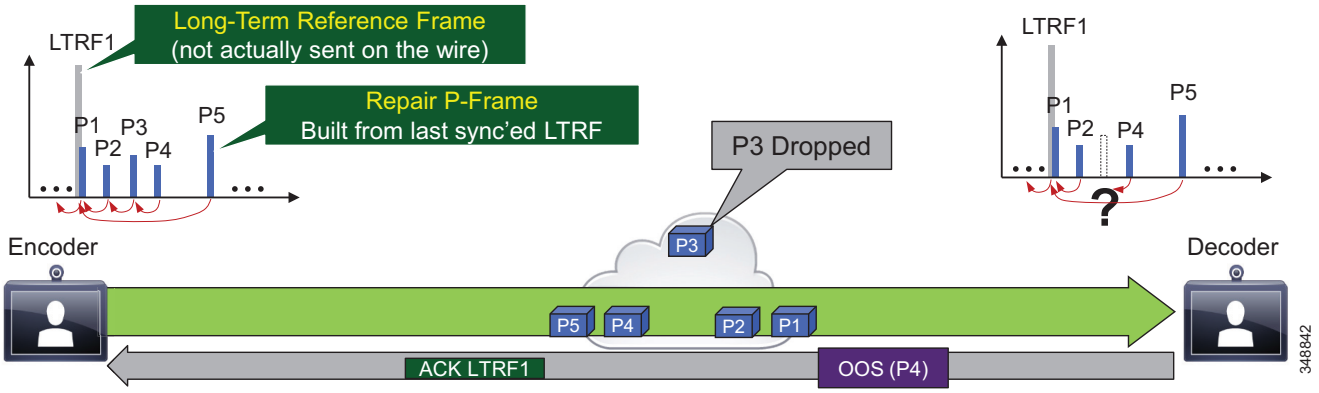
Long Term Reference Frame (LTRF)

A Long Term Reference Frame (LTRF) is a reference frame that is stored in the encoder and decoder until they receive an explicit signal to do otherwise. (Up to 15 LTRFs are supported by H.264.) Typically (without LTRF) an intra-frame is used for encoder/decoder resynchronization after packet loss.

LTRFs can provide benefits over normal infra-frames as an alternative method for encoder/decoder resynchronization. Typically, the encoder inserts LTRFs periodically and at the same time instructs the decoder to store one or more of those LTRFs (see Figure 13-9).

A repair P-frame uses a previous LTRF that has been decoded correctly as a reference. The repair P-frame is used in response to a missing frame or its reference frame. Because the acknowledged LTRF is known to have been correctly received at the decoder, the decoder is known to be back in-sync if it can correctly decode a repair P-frame.

Figure 13-9 Long Term Reference Frame (LTRF)

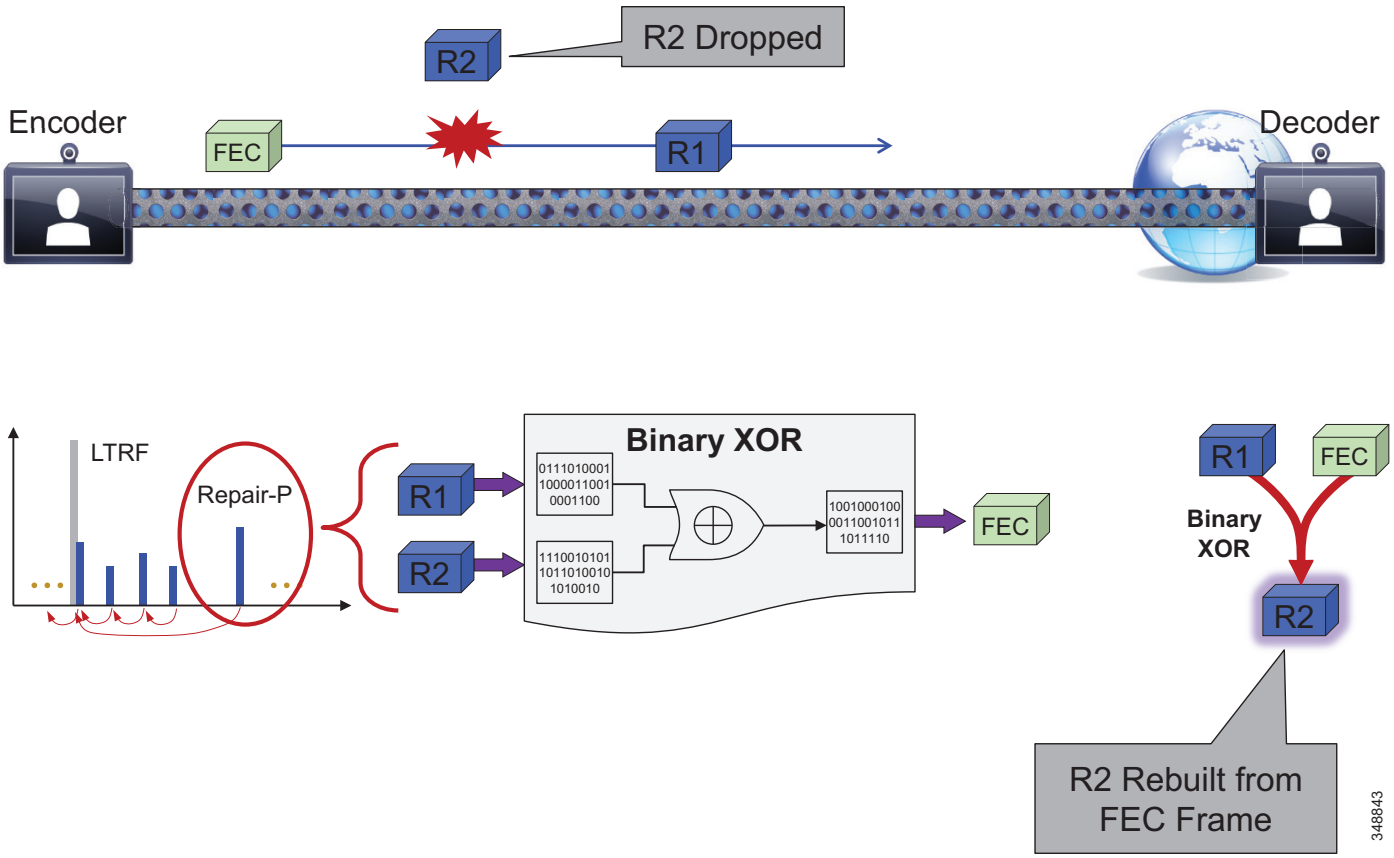


As Figure 13-9 illustrates, LTRFs keep the encoder and decoder in sync with active feedback messages. The encoder instructs the decoder to store raw frames at specific sync points as Long Term Reference Frames (part of the H.264 standard), and the decoder uses "back channel" (RTCP) to acknowledge the LTRFs. When a frame is lost, the encoder creates a Repair P-frame based on the last synchronized LTRF instead of generating a new I-frame, thus saving bandwidth.

Forward Error Correction (FEC)

Forward error correction (FEC) provides redundancy to the transmitted information by using a predetermined algorithm (see Figure 13-10). The redundancy allows the receiver to detect and correct a limited number of errors occurring anywhere in the message, without the need to ask the sender for additional data. FEC gives the receiver an ability to correct errors without needing a reverse channel to request retransmission of data, but this advantage is at the cost of a fixed higher forward channel bandwidth. FEC protects the most important data (typically the repair P-frames) to make sure the receiver is receiving those frames. The endpoints do not use FEC on bandwidths lower than 768 kbps, and there must also be at least 1.5% packet loss before FEC is introduced. Endpoints typically monitor the effectiveness of FEC, and if FEC is not efficient, they make a decision not to do FEC.

Figure 13-10 Forward Error Correction (FEC)

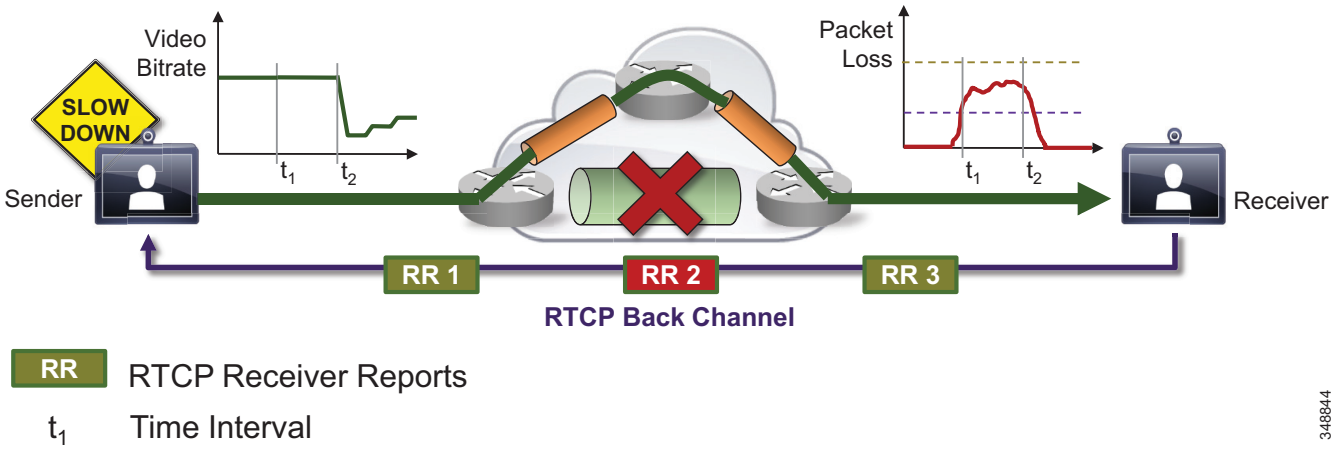


As Figure 13-10 illustrates, FEC enables the decoder to recover from a limited amount of packet loss without losing synchronization. It can be applied at different levels (for example, *X* FEC packets every *N* data packets) to protect "important" frames in lossy environments. The correction code can be basic (binary XOR) or more advanced (Reed-Solomon). The trade-off is increased bandwidth usage, therefore it is best suited for non-bursty loss.

Rate Adaptation

Rate adaptation, or dynamic bit rate adjustments, adapt the call rate to the variable bandwidth available, down-speeding or up-speeding the video bit rate based on the packet loss condition (see Figure 13-11). Once the packet loss has decreased, up-speeding will occur. Some endpoints use a proactive sender-initiated approach by utilizing RTCP. In this case the sender is constantly reviewing the RTCP receiver reports and adjusting its bit rate accordingly. Other endpoints use a receiver-initiated approach, adjusting via call signaling (H.323 flow control, TMBRR, SIP Re-invite) or an explicit request in the RTCP messages.

Figure 13-11 Rate Adaption



As illustrated in Figure 13-11, the receiver observes delay and packet loss over periods of time and signals back using RTCP Receiver Reports (RR). the reports cause the sender to adjust its bit rate to adapt to network conditions (down-speeding or up-speeding of bit rate).

Two approaches are possible with rate adaption:

- Sender-initiated adjustment based on RTCP Receiver Reports
- Receiver-initiated adjustment via call signaling (H.323 flow control, TMBRR, SIP Re-invite) or explicit request in RTCP message

Summary

- Burstiness of traffic and mobility of the endpoints make deterministic provisioning for interactive video difficult for network administrators.
- Media resiliency mechanisms help mitigate the impact of video traffic on the network and the impact of network impairments on video. (See Table 13-6.)
- Dynamic rate adaptation creates an opportunity for more flexible provisioning models for interactive video in enterprise networks.
- Media resiliency and rate adaptation also help preserve the user experience when video traffic traverses the Internet or non-QoS-enabled networks.

Table 13-6 Media Resilience Support in Cisco Collaboration Video Endpoints

Endpoint or Bridge	Encoder Pacing	Rate Adaption	FEC	LTRF Repair
8800 Series	Yes	No	No	No
9900 Series	No	No	No	No
DX Series	Yes	Yes	No	No
WebEx	Yes	Yes	Yes	No
TX Series	Yes	Yes	No	Yes
Jabber	Yes	Yes	Yes	Yes
C, EX, MX, SX, and Profile Series	Yes	Yes	Yes	Yes
TelePresence Server	Yes	Yes	Yes	Yes
MCU	Yes	Yes	Yes	Yes
Cisco Meeting Server	Yes	Yes	Yes	Yes

QoS Architecture for Collaboration

Quality of Service (QoS) ensures reliable, high-quality voice and video by reducing delay, packet loss, and jitter for media endpoints and applications. QoS provides a foundational network infrastructure technology that is required to support the transparent convergence of voice, video, and data networks. With the increasing amount of interactive applications (particularly voice, video, and immersive applications), real-time services are often required from the network. Because these resources are finite, they must be managed efficiently and effectively. If the number of flows contending for such priority resources were not limited, then as these resources become oversubscribed, the quality of all real-time traffic flows would degrade, eventually to the point of become useless. "Smart" media techniques, QoS, and admission control ensure that real-time applications and their related media do not over-subscribe the network and the bandwidth provisioned for those applications. These smart media techniques coupled with QoS and, where needed, admission control, can be a powerful set of tools to protect real-time media from non-real-time network traffic and protect the network from over-subscription and the potential loss of quality of experience for all voice and video applications.

Admission control and QoS are complementary. Admission control requires QoS, but QoS may be deployed without admission control. Later in this chapter, admission control and its relationship to QoS are discussed further.

Figure 13-12 illustrates the approach to QoS used in this chapter. This approach consists of the following phases, discussed further in the sections that follow:

- Identification and Classification, page 13-17

  This phase involves the concepts of trust and techniques for identifying media and signaling for trusted and untrusted endpoints. It includes the process of mapping the identified traffic to the correct DSCP to provide the media and signaling with the correct per-hop behavior end-to-end across the network for both trusted and untrusted endpoints.

- WAN Queuing and Scheduling, page 13-32

  This phase consists of general WAN queuing and scheduling, the various types of queues, and recommendations for ensuring that collaboration media and signaling are correctly queued on egress to the WAN.

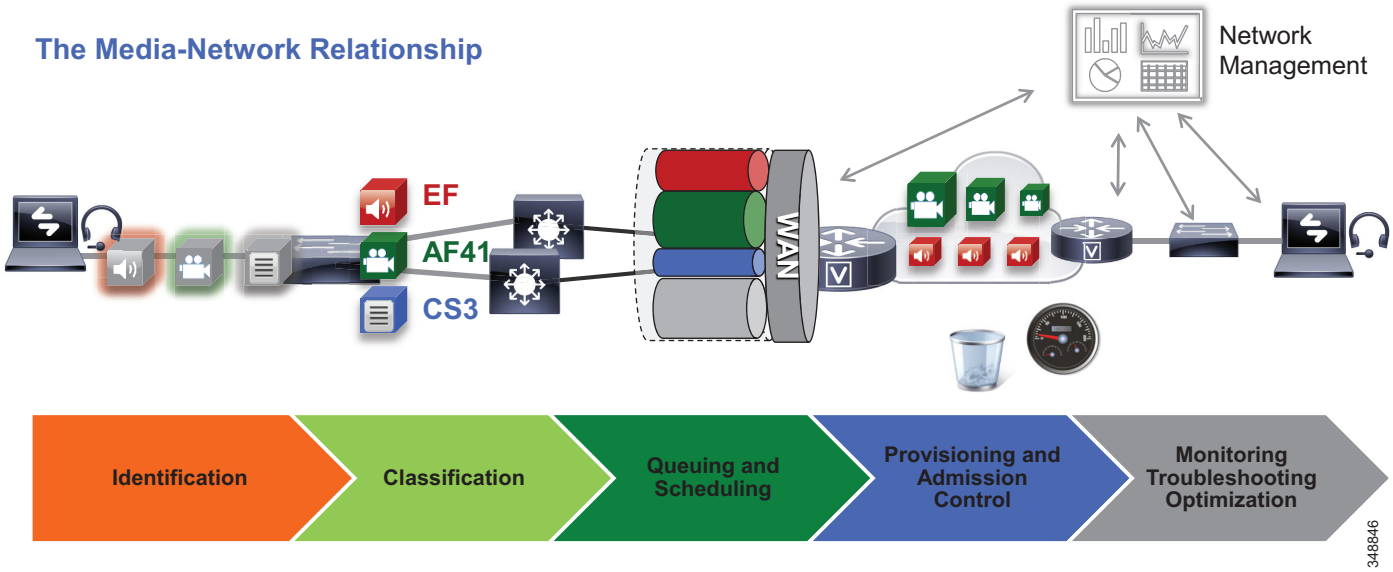
- Provisioning and Admission Control, page 13-37

  This phase involves provisioning of bandwidth in the network and determining the maximum bit rate that groups of endpoints will utilize. This is also where call admission control can be implemented in areas of the network where it is required.

- Monitoring, Troubleshooting, and Optimization

  This phase is crucial to the proper operation and management of voice and video across the network; however, it is not discussed in this chapter. For information on these tasks, see the chapter on Network Management, page 27-1.

Figure 13-12 Elements of the QoS Architecture for Collaboration



Identification and Classification

This section discusses the concepts of trust and techniques for identifying media and signaling for trusted and untrusted endpoints. It includes the process of mapping the identified traffic to the correct DSCP to provide the media and signaling with the correct per-hop behavior end-to-end across the network for both trusted and untrusted endpoints.

QoS Trust and Enforcement

The enforcement of QoS is crucial to any real-time audio, video, or immersive video experience. Without the proper QoS treatment (classification, prioritization, and queuing) through the network, real-time media can potentially incur excessive delay or packet loss, which compromises the quality of the real-time media flow. In the QoS enforcement paradigm, the issue of trust and the trust boundary is equally important. Trust refers to the endpoint or device permitting or "trusting" the QoS marking (Layer 2 CoS or Layer 3 IP DSCP) of the traffic and allowing it to continue through the network. The trust boundary is the place in the network where the trust occurs. It can occur at any place in the network, but we recommend enforcing trust at the network edge, such as the LAN access ingress or the WAN edge or both if feasible and applicable. The WAN edge is another area of traffic ingress, and sometimes

service providers re-mark traffic for usage throughout their network (service provider network). Because of this situation, it is important to re-mark the traffic back to the appropriate values to ensure continuity through the enterprise network end-to-end.

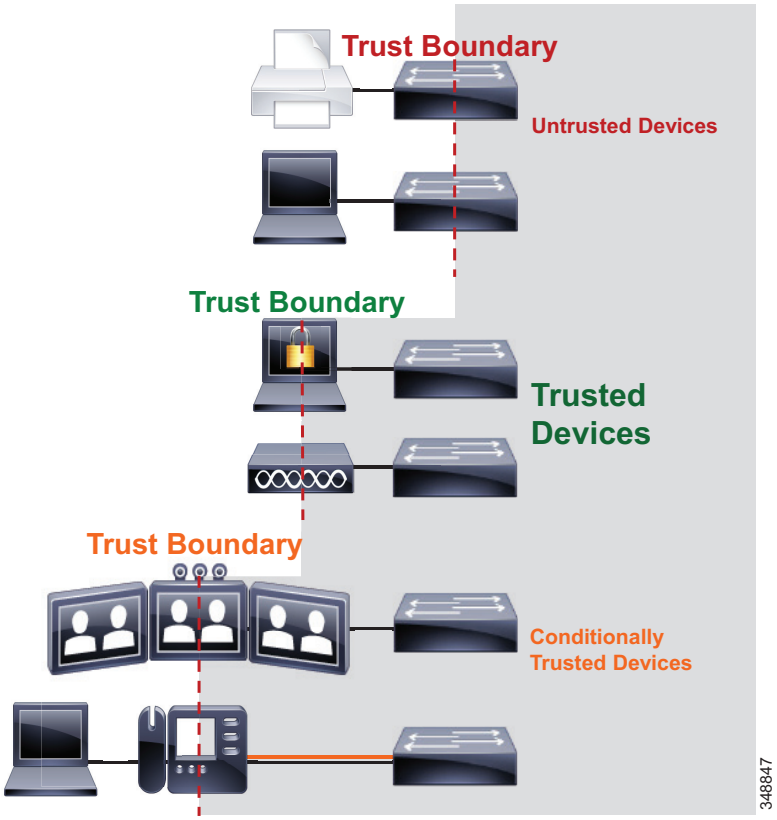
In a Cisco converged network with Cisco IP phones and video endpoints, switches can be configured to detect the phones using Cisco Discovery Protocol (CDP), and the switch can then trust the Differentiated Services Code Point (DSCP) marking of packets that the Cisco IP phones and video endpoints send without trusting the markings of the PC connected to the switch port of the IP phone or video endpoint. This is referred to as conditional trust and is commonplace in protected VLANs where only Cisco IP phones are admitted (referred to as voice VLANs) and where their packet marking is trusted by the switches and passed through the network unchanged. Administrators generally do not trust the traffic that comes from VLANs where untrusted clients (such as PCs or Macs) are typically located (referred to as data VLANs). The packets that come from devices in the data VLAN or equivalent areas of the network typically get remarked to best effort (IP DSCP 0).

From a trust perspective, there are three main categories of endpoints:

- **Untrusted endpoints** — Unsecure PCs, Macs, or hand-held mobile devices
- **Trusted endpoints** — Secure PCs and servers, video conferencing endpoints, access points, analog and video conferencing gateways, and other similar devices where CDP is not available
- **Conditionally trusted endpoints** — Cisco IP phones as well as Cisco TelePresence endpoints that support CDP

Figure 13-13 illustrates these three types of devices.

Figure 13-13 Trust Boundaries



The trust boundary should be set as close to the endpoints as technically and administratively feasible. The recommendation is to set trust on the switch and use voice VLANs for collaboration media and signaling, and use data VLANs for non-collaboration data traffic. See the section on Campus Access Layer, page 3-4, for more information on Layer 2 access design.

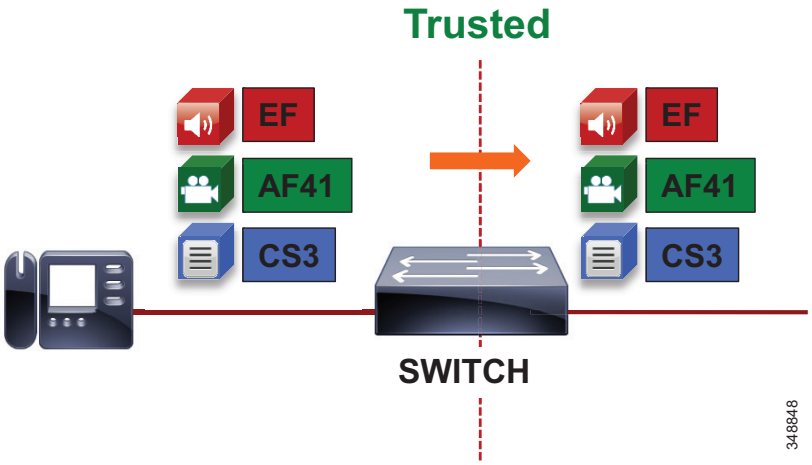
Classification and Marking

Once the trust boundary is established, QoS enforcement can be put into place for two categories of devices: trusted and untrusted. This section discuss classification and marking for trusted and untrusted devices.

Trusted Endpoints

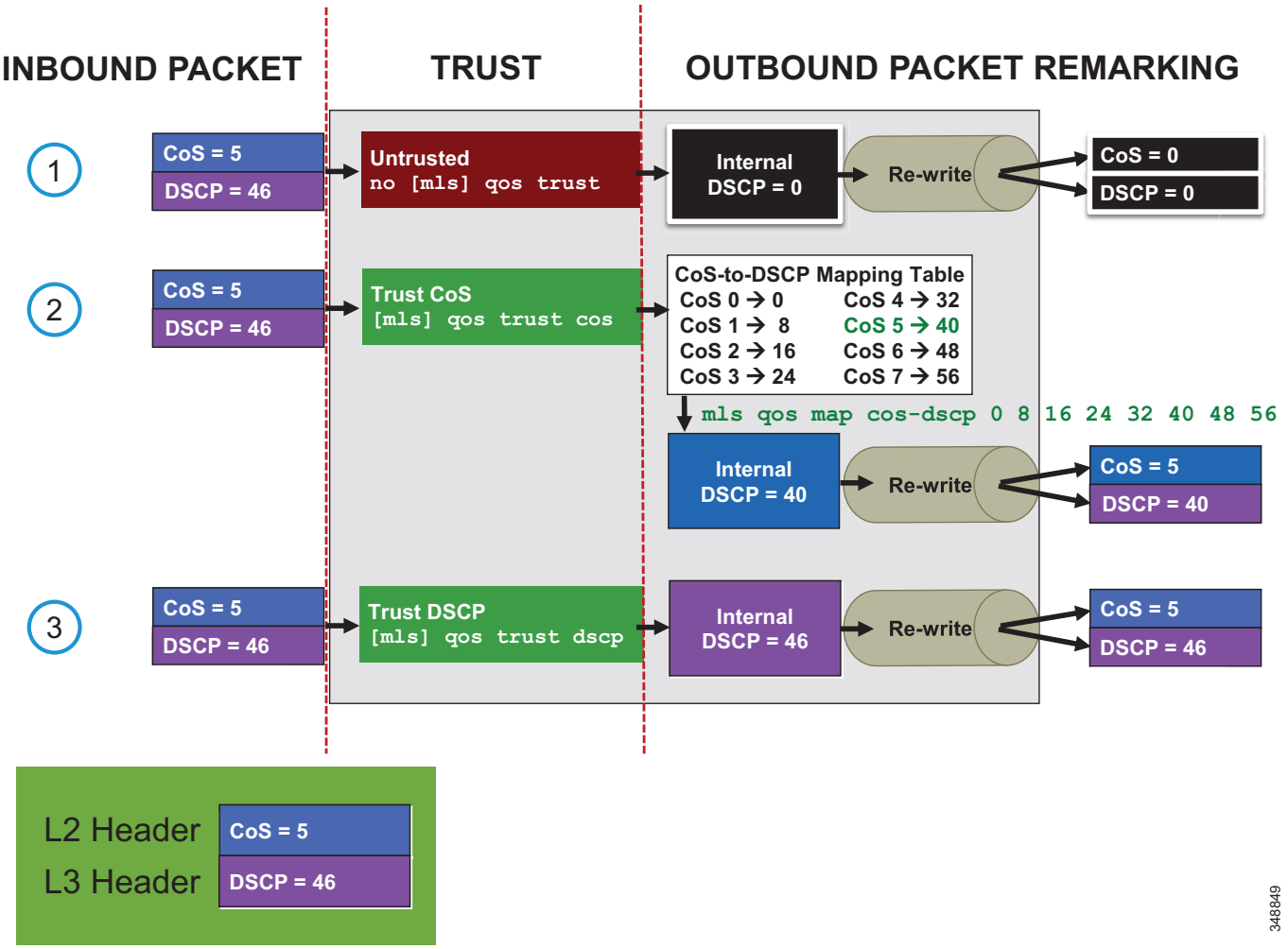
For trusted and conditionally trusted endpoints, the DSCP marking of packets on ingress into the switch are trusted and rewritten to the same value on egress. Figure 13-14 illustrates marking of audio, video, and signaling traffic for trusted endpoints, and the switch trusting these markings.

Figure 13-14 Trusted Endpoint Re-marking



For Cisco switches configured with trusted or conditionally trusted ports, the switch either uses CoS to map to DSCP or it uses the original DSCP and maps it to the outbound packet IP header DSCP. Figure 13-15 illustrates the inbound packet marking at Layer 2 (CoS) and Layer 3 (DSCP); the type of trust – trusted (CoS Trust or DSCP Trust) or untrusted; and the internal switch packet rewriting process based on CoS trust or DSCP trust.

Figure 13-15 Inbound and Outbound Switch Packet Marking



Multi-Layer Switching (MLS) commands are used in Figure 13-15 as an example only. MLS platforms include the Cisco 2960, 3560, and 3750 Series switch platforms. On all other currently shipping switch platforms (including the Cisco 3650, 3850, 4500, 6500, and 6800 Series switch platforms) trust is enabled by default.

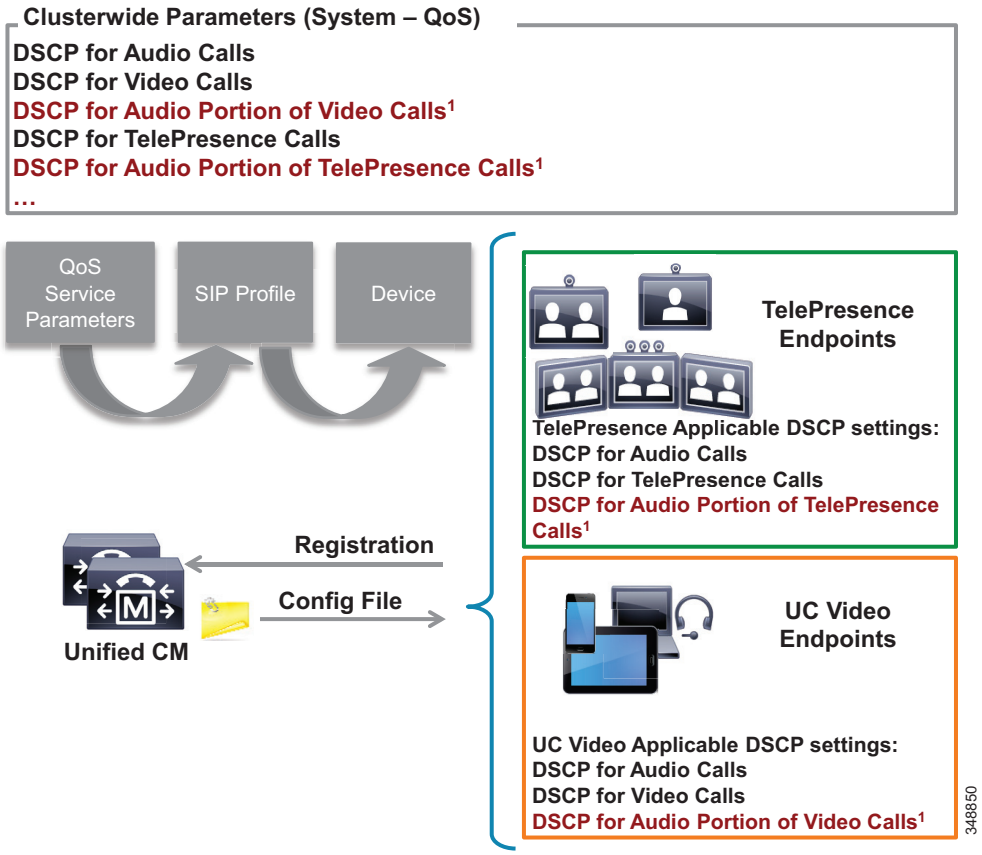
Figure 13-15 shows three events:

1. A packet marked CoS 5 and DSCP 46 comes inbound on an untrusted port. An internal DSCP of 0 (BE) is used to rewrite the outbound packet CoS and DSCP to 0.
2. A packet marked CoS 5 and DSCP 46 comes inbound on a trusted port (CoS trust). A lookup is done on a CoS-to-DSCP mapping table to map CoS 5 to an internal DSCP of 40. An internal DSCP of 40 is used to rewrite the outbound packet CoS to 5 and DSCP to 40. Note that the CoS-to-DSCP map table has defaults but can be modified to any static CoS-to-DSCP mapping. For example, CoS 5 could be mapped to DSCP 46 (EF).
3. A packet marked CoS 5 and DSCP 46 comes inbound on a trusted port (DSCP trust). An internal DSCP of 46 (EF) is used to rewrite the outbound packet CoS to 5 and DSCP to 46 (EF).

For CDP-capable Cisco IP Phones, Cisco CTS, Cisco IP Video Surveillance cameras, and Cisco Digital Media Players (as opposed to software clients such as Jabber), we recommend using the CDP conditional trust and passing the marking of the trusted endpoint through the network. When electing to trust Cisco IP Phones, you must trust CoS because the phones can re-mark only PC traffic at Layer 2. Trusted endpoints derive their DSCP marking from Unified CM. DSCP for endpoints is configured in the Unified CM service parameters under **Clusterwide Parameters (System - QoS)**.

Unified CM houses the QoS configuration for endpoints in two places: in the service parameters for the CallManager service and in the SIP Profile applicable only to SIP devices. The SIP Profile configuration of QoS settings overrides the service parameter configuration. This allows the Unified CM administrator to set different QoS policies for groups of endpoints (see Bandwidth Management Design Examples, page 13-91). During endpoint registration, Unified CM passes this QoS configuration to the endpoints in a configuration file over TFTP. This configuration file contains the QoS parameters as well as a number of other endpoint specific parameters. For QoS purposes there are two categories of video endpoints: TelePresence endpoints (any endpoint with TelePresence in the phone type name) and all other non-TelePresence video endpoints referred to as "UC Video Endpoints" in this document. Figure 13-16 illustrates how the two categories of Cisco video endpoints derive DSCP. Keep in mind that these categories apply only to QoS and call admission control (see the section on Enhanced Location CAC for TelePresence Immersive Video, page 13-59).

Figure 13-16 How Cisco Endpoints Derive DSCP



The parameters **DSCP for Audio Portion of Video Calls** and **DSCP for Audio Portion of TelePresence Calls**, shown in Figure 13-16, currently are not supported on all video endpoints. See Table 13-8 for information on which endpoint types support these parameters.

The configuration file is populated with the QoS parameters from the CallManager service parameters or the SIP Profile, when configured, and sent to the endpoint upon registration. The endpoint then uses the correct DSCP parameters for each type of media stream, depending on which category of endpoint it is. Table 13-7 lists the DSCP parameter, the type of endpoint, and the type of call flow determining the DSCP marking of the stream.

Table 13-7 DSCP for Basic Call Flows[1]

DSCP Parameter	TelePresence Endpoint	UC Video Endpoint	Call Flow
DSCP for Audio Calls	Yes[2]	Yes	Voice only
DSCP for Video Calls	N/A	Yes	Video – Audio and video stream of a video call, unless the endpoint supports the **DSCP for Audio Portion of Video Calls** parameter
DSCP for Audio Portion of Video Calls[3]	N/A	Yes	Audio stream of a video call – Applicable only to endpoints that support this parameter
DSCP for TelePresence Calls	Yes	N/A	Immersive video – Audio and video of an immersive video call, unless the endpoint supports the **DSCP for Audio Portion of TelePresence Calls** parameter.
DSCP for Audio Portion of TelePresence Calls[3]	Yes	N/A	Audio stream of a video call – Applicable only to endpoints that support this parameter

1. The DSCP settings for Multi-Level Priority and Preemption (MLPP) are not discussed here. For more information about MLPP and QoS settings, refer to the latest version of the System Configuration Guide for Cisco Unified Communications Manager.
2. Of the TelePresence video endpoints, only the DX Series endpoints support **DSCP for Audio Calls** in addition to the corresponding **DSCP for Audio Portion of TelePresence Calls**, and only if the endpoints are running CE Software.
3. This parameter is not currently supported on all video endpoints. See Table 13-8 for information on which endpoint types support this parameter.

Table 13-8 Endpoint Support for DSCP Parameters for the Audio Portion of Video and TelePresence Calls

Video Endpoint	DSCP for Audio Portion of Video Calls	DSCP for Audio Portion of TelePresence Calls
8800 Series	Yes	N/A
8900 Series	No	N/A
9900 Series	No	N/A
Jabber	Yes[1]	No
DX650; DX70 and DX80 with non-CE Software	Yes	Yes[2]
TX Series	N/A	Yes

Table 13-8 Endpoint Support for DSCP Parameters for the Audio Portion of Video and TelePresence Calls (continued)

Video Endpoint	DSCP for Audio Portion of Video Calls	DSCP for Audio Portion of TelePresence Calls
IX Series	N/A	No
CE 8.x Software Series (DX70, DX80, SX Series, MX Series G2, MX700, MX800)	N/A	Yes
TC 7.1.4 Software Series (C Series, Profile Series, EX Series, MX Series G1)	N/A	Yes
EX Series (TC Software)	N/A	Yes

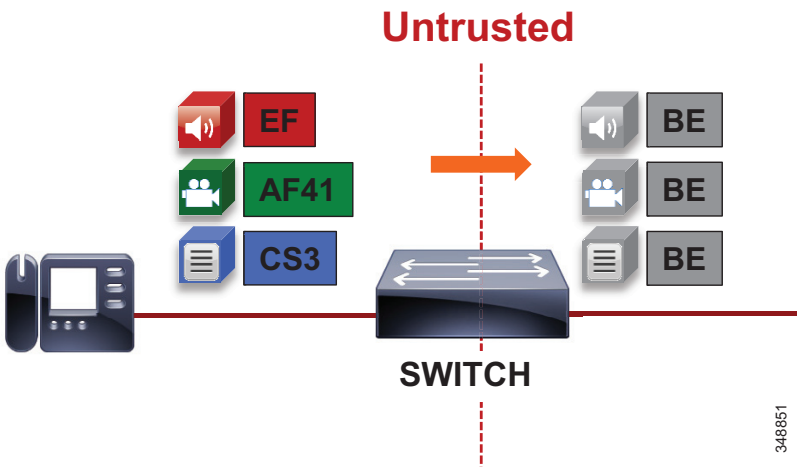
1. Jabber for Windows uses Group Policy Objects to mark traffic on the PC. All other Jabber clients are able to mark DSCP natively.
2. To enable the DX70 and DX80 to use DSCP for TelePresence calls as well as DSCP for the audio portion of TelePresence calls, you must upgrade to CE Software.

Due to these new features and system-wide capabilities, the current DSCP defaults are not always the recommended values. This is discussed in further detail in the Bandwidth Management Design Examples, page 13-91.

Untrusted Endpoints and Clients

For untrusted endpoints the DSCP marking of packets on ingress into the switch is untrusted and rewritten to 0 (BE). Figure 13-17 illustrates untrusted endpoints marking audio, video, and signaling traffic, and the switch rewriting this value on the outbound packet.

Figure 13-17 Untrusted Endpoint Re-marking



In general, trusting markings that can be set by users on their PCs, Macs, or hand-held mobile devices is not recommended. Users can abuse provisioned QoS policies if permitted to mark their own traffic (have administrative control of the OS). For example, if a DSCP of EF has been provisioned over the network, a PC user can configure all their traffic to be marked to EF, which will hijack network priority queues to service non-real-time traffic. Such abuse could easily ruin the service quality of real-time applications throughout the enterprise. On the other hand, if enterprise controls are in place that centrally

administer PC QoS markings, such as Global Policy Objects in Windows environments, then it may be possible to trust the PC markings. For Macs running OSX and hand-held mobile clients, the question remains whether to trust the markings from them or not. This method is covered in more detail in the section "Utilizing the Operating System for QoS Trust, Classification and Marking". The general rule is not to trust any of these personal computing devices, and a method for re-marking traffic is required.

A different method from trust is required to ensure that the media and signaling streams from the software clients such as Jabber are able to get classified and marked appropriately. One method consists of mapping identifiable media and signaling streams based on specific protocol ports, such as UDP and TCP ports, then making use of network access lists to remark QoS of the signaling and media streams based on those protocol port ranges. This method applies to all Cisco Jabber clients because they all behave similarly when allocating media and signaling port ranges. This method ranges from using the network to create policies based on access lists to accomplish packet DSCP remarking, to using the Windows OS itself (Jabber for Windows clients only apply here) and then trusting the marking from the PC in the network.

This method is the most widely deployed and recommended method to achieve QoS with Cisco Jabber clients simply because of the trust issue. The Jabber clients are Cisco Jabber for Windows, Cisco Jabber for Mac OS, Cisco Jabber for iPhone, Cisco Jabber for iPad and Cisco Jabber for Android.

The concept is simple. As all of the traffic from the PC cannot be trusted, an access list is used in the network access layer equipment to identify the media and signaling streams based on UDP port ranges and to re-mark them to appropriate values. Although this technique is easy to implement and can be widely deployed, it is not a secure method.

Figure 13-18 illustrates using network access control lists (ACLs) to map identifiable media and signaling streams to DSCP.

Figure 13-18 Mapping UDP/TCP Port Ranges to DSCP

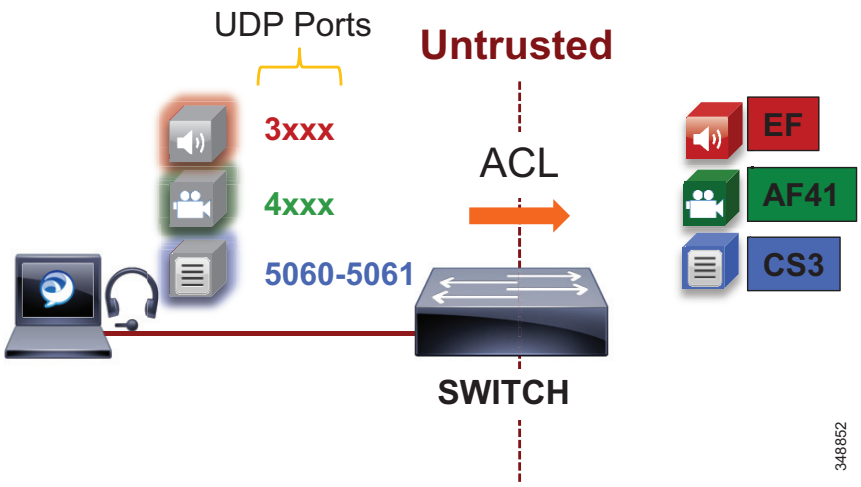


Figure 13-18 illustrates the following example ACL-based QoS policy for Jabber clients:

- UDP Port Range 3xxx Mark to DSCP EF
- UDP Port Range 4xxx Mark to DSCP AF41
- TCP Port 5060-5061 Mark to DSCP CS3

Note The following example access control list is based on the Cisco Common Classification Policy Language (C3PL). Refer to your specific switch or router configuration guides to achieve the same policy on a Cisco device that does not support C3PL or for any updated commands in C3PL. This configuration is portable to all currently shipping switches including Modular QoS CLI-MQC, Multi-Layer Switching (MLS), and C3PL.

```
! This section configures the ACLs to match the UDP Port ranges
   access-list 100 permit udp any range 3000 3999 any
   access-list 101 permit udp any range 4000 4999 any
   access-list 102 permit tcp any range 5060 5061 any

! This section configures the classes that match on the ACL's
   class-map JABBER-VOICE
     match access-group 100
   class-map JABBER-VIDEO
     match access-group 101
   class-map JABBER-SIP
     match access-group 102

! This section configures the policy-map matching the classes configured above and setting
DSCP for JABBER Voice, Video and SIP Signaling on ingress (Generic default DSCP values are
used; see design considerations for recommended values for Jabber).
policy-map INGRESS-MARKING
   class JABBER-VOICE
     set dscp ef
   class JABBER-VIDEO
     set dscp af41
   class JABBER-SIP
     set dscp cs3
   class class-default

! This section applies the policy-map to the Interface
  Switch (config-if)# service-policy input INGRESS-MARKING
```

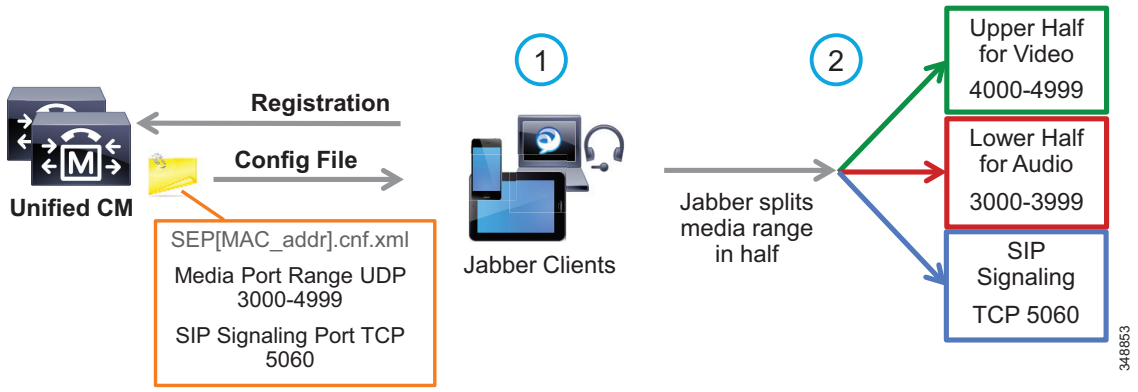
QoS for Cisco Jabber Clients

As discussed, this method involves classifying media and signaling by identifying the various streams from the Jabber client based on IP address, protocol, and/or protocol port range. Once identified, the signaling and media streams can be classified and remarked with a corresponding DSCP. The protocol port ranges are configured in Unified CM and are passed to the endpoint to use during device registration. The network can then be configured via access control lists (ACLs) to classify traffic based on IP address, protocol, and protocol port range and then re-mark the classified traffic with the appropriate DSCP as discussed above.

Cisco Jabber provides identifiable media streams based on UDP protocol port ranges and identifiable signaling streams based on TCP protocol port ranges. In Unified CM, the signaling port for endpoints is configured in the SIP Security Profile, while the media port range is configured in the SIP Profile of the Cisco Unified CM administration pages.

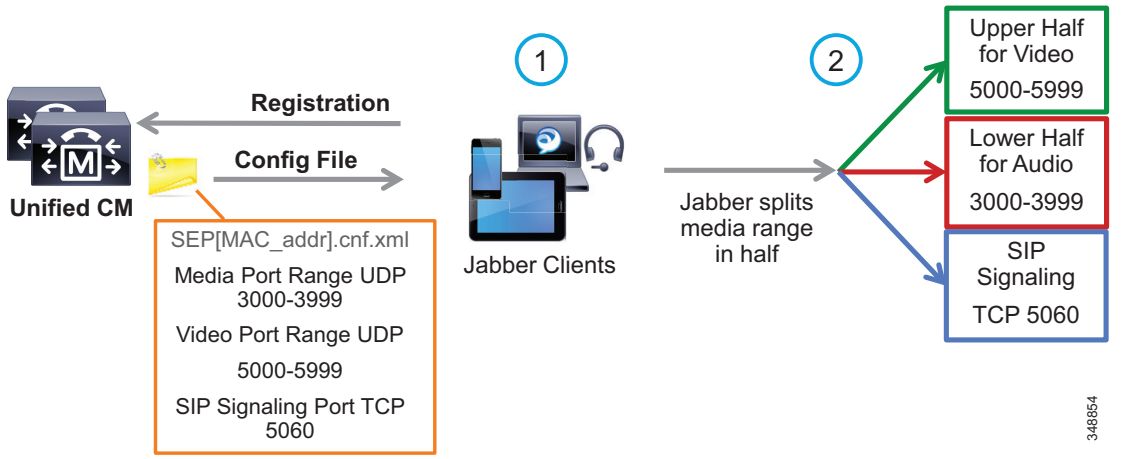
For the media port range, all endpoints and clients use the SIP profile parameter **Media Port Range** to derive the UDP ports used for media. By default media port ranges are configured with **Common Port Range for Audio and Video**. When Jabber clients receive this port range in their Config file, they split the port range in half and use the lower half for the audio streams of both voice and video calls and the upper half for the video streams of video calls. Jabber does not place the audio of a video call in the video UDP port range when using the **Media Port Range > Common Port Range for Audio and Video** configuration. This is illustrated in Figure 13-19.

Figure 13-19 Media and Signaling Port Range – Common



Jabber can also use the **Media Port Range > Separate Port Range for Audio and Video** configuration. In this configuration the Unified CM administrator can configure a non-contiguous audio and video port range as illustrated in Figure 13-20.

Figure 13-20 Media and Signaling Port Range – Separate



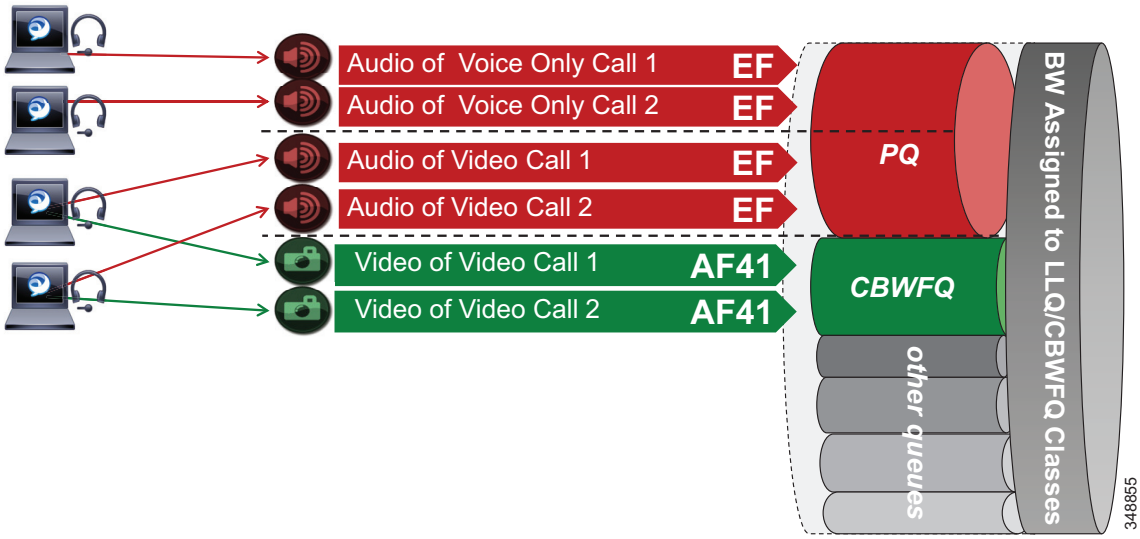
Due to the behavior of Jabber clients regarding UDP port range assignment, it is often not possible to map Enhanced Locations Call Admission Control (EL-CAC) bandwidth deductions correctly with QoS markings. CAC deducts bandwidth for audio-only calls out of the voice pool, while both audio and video bandwidth of a video call is deducted out of the video pool. To be consistent with the admission control logic, audio streams of voice-only calls would need to be marked as EF while both audio and video streams of video calls would need to be marked AF41. The differentiation of audio between audio of

voice-only calls and audio of video calls is not possible when using Cisco Jabber clients and UDP port ranges to map identifiable media streams. As a result, this technique is effective to achieve QoS only. Therefore, we recommend over-provisioning the priority queue for EF traffic to account for the audio of video sessions from Jabber clients that will send audio as EF, or using an alternate DSCP. Some strategies are discussed in the Bandwidth Management Design Examples, page 13-91.

Caution **Security Alert**: By utilizing identifiable media streams for QoS classification at the network level, the trust model does *not* extend to the application itself. Apart from prioritizing streams from the intended application, other applications "could" potentially be configured to use the same identification criteria (media port range), and therefore achieve network prioritization. Because this unintended traffic would not be accounted for in CAC or in the provisioning of the network, severe overall impacts to real-time conversations can occur. It is good practice to define restricted port ranges to identify media streams when possible.

When utilizing this technique, it is important to ensure that the audio portion of these video calls that will be re-marked to the audio traffic class (EF) and the video portions re-marked to the video traffic class (AF4) are provisioned in the network accordingly. Figure 13-21 is an example of placing audio traffic into a Priority Queue (PQ) and video traffic into a Class Based Weighted Fair-Queue (CBWFQ). Note that, because it is not possible to differentiate the audio from voice-only calls versus the audio from video calls with port ranges in Cisco Jabber endpoints, all audio using this technique will be re-marked to EF. It is important to provision the PQ adequately to support voice-only and the audio portion of video calls. An example of such provisioning is illustrated in Figure 13-21. For more information on the design and deployment recommendations for provisioning queuing and scheduling in the network, see the section on WAN Queuing and Scheduling, page 13-32.

Figure 13-21 Provisioning Jabber QoS in the Network



According to RFC 3551, when RTCP is enabled on the endpoints, it uses the next higher odd port. For example, a device that establishes an RTP stream on port 3500 would send RTCP for that same stream on port 3501. This function of RTCP is also true with all Jabber clients. RTCP is common in most call

flows and is commonly used for statistical information about the streams and to synchronize audio and video in video calls to ensure proper lip-sync. In most cases, video and RTCP can be enabled or disabled on the endpoint itself or in the common phone profile settings.

Utilizing the Network for Classification and Marking

Based on the identifiable media and signaling streams created by the Jabber client, common network QoS tools can be utilized to create traffic classes and re-mark packets according to these classes.

These QoS mechanisms can be applied at different layers, such as the access layer (access switch), which is closest to the endpoint and the router level in the distribution, core, or services WAN edge. Regardless of where classification and re-marking occur, we recommend using DSCP to ensure end-to-end per-hop behaviors.

As previously mentioned, Cisco Unified CM allows the port range utilized by SIP endpoints to be configured in the SIP profile. As a general rule, a port range of a minimum of 100 ports (for example, 3000 to 3099) is sufficient for most scenarios. A smaller range could be configured, as long as there are enough ports for the various audio, video, and associated RTCP ports (RTCP runs over the odd ports in the range).

Note When deploying Jabber clients in networks where SCCP voice-only endpoints are deployed, the SCCP endpoints use a non-configurable hard-coded range of 16384 to 32767 for voice-only calls. Due to this, SCCP voice-only calls could run over the same range as SIP video-enabled endpoint calls if you do not change the media port range for SIP devices. If you are deploying a collaboration solution with endpoints that are configured to use SCCP, then we recommend setting the media port range of Jabber clients outside of the 16384 to 32767 range. The above examples of 3000 to 4999 for video-enabled Jabber clients and 3000 to 3999 for voice-only Jabber clients work very well to avoid overlap with SCCP endpoints.

The recommendation to avoid overlap applies to other SIP-based video endpoints as well. To avoid overlap with SCCP-based audio endpoint ranges, the SIP-based video endpoints should also be allocated a port range that does not overlap with SCCP-based audio port range (16384 to 32767) or the Jabber clients' media port range.

Access Layer (Layer 2 Definitions)

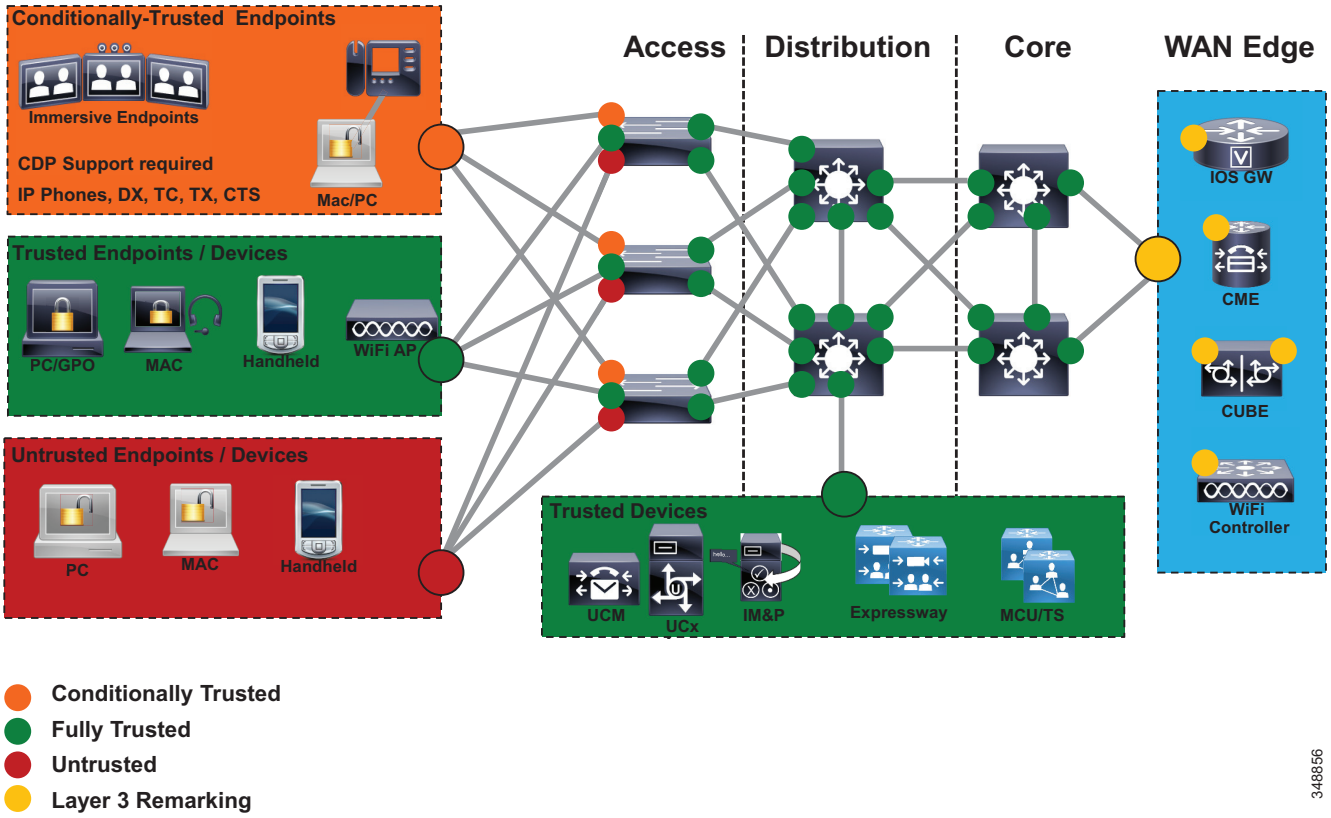
When using the access layer to classify traffic, the classification occurs at the ingress of traffic into the network, allowing the flows to be identified as they enter. In environments where QoS policies are applied not only in the WAN but also within the LAN, all upstream components can rely on traffic markings when processing. Classification at the ingress allows different methods to be utilized based on different types of endpoints. Physical endpoints such as IP phones can rely on mechanisms such as the Cisco Discovery Protocol (CDP) or Link Layer Discovery Protocol-Media Endpoint Discovery (LLDP-MED) to establish a trust relationship. Once the device is identified as trusted, QoS markings received from the device are trusted throughout the network.

Configuring QoS policies in the access layer of the network could result in a significant number of devices that require configuration, which can create additional operational overhead. The QoS policy configurations should be standardized across the various switches of the access layer through templates. You can use configuration deployment tools to relieve the burden of manual configuration.

Distribution, Core, and Services WAN Edge (Layer 3 Definitions)

Another location where QoS marking can take place is at the Layer 3 routed boundary. In a campus network, Layer 3 could be in the access, distribution, core, or services WAN edge layer. The recommendation is to build the trust boundary and classify and re-mark at the access. Then trust through the distribution and core of the network, and finally re-classify and re-mark at the WAN edge. For smaller networks such as branch offices where there no Layer 3 switching components are deployed, QoS marking can be applied at the WAN edge router. At Layer 3, QoS policies are applied to the Layer 3 routing interfaces. In most campus networks these would be VLAN interfaces, but they could also be Fast Ethernet or Gigabit Ethernet interfaces. Figure 13-22 illustrates the areas of the network where the various types of trust are applied in relation to the places in the network – access, distribution, core, and WAN Edge.

Figure 13-22 Trust and Enforcement – Places in the Network



Utilizing the Operating System for QoS Trust, Classification, and Marking

Another method of QoS trust for Cisco Jabber clients is to allow the operating system on which the applications run to mark the QoS of the media and signaling at the request of the application. The benefit of this method is that it allows the network operators to extend the QoS trust model to the operating system itself, and then they can configure the network to "trust" the QoS markings and pass them through the network. It is not a common enterprise practice to extend QoS trust to the Windows PCs, Mac OS, and hand-held devices. The reason for this is that this method trusts all traffic from the device, not just

traffic from authenticated application communication. These applications can be installed and used on these devices to "hijack" a priority QoS and defeat the original purpose of deploying QoS in the first place. Through administrative global policies, administrators can manage some operating systems such as Windows OS or user access controls to ensure that the OS does not accept unwanted applications or configurations. In these cases, it might be acceptable to use this method of QoS trust.

On Windows 7 and 8 operating systems it is necessary to configure specific policies, while in Mac OS, Apple iOS, and Android devices the OS natively marks at the request of the application without any specific configuration necessary.

The following sections discuss the Cisco Jabber clients and describe how each operating system functions with regard to application QoS classification and marking. Everything described in these sections relates to Layer 3 DSCP marking and not Layer 2 Class of Service (CoS):

- Classification in Windows 7 and 8, page 13-30
- Classification in Mac OS, page 13-31
- Classification in Apple iOS (iPhone and iPad), page 13-32
- Classification in Android, page 13-32

Classification in Windows 7 and 8

Microsoft Windows 7 and Windows 8 take a different approach when it comes to QoS marking by the operating system because Microsoft's security enhancement, User Account Control (UAC), does not allow a regular application to set DSCP markings on IP packets, which is considered to be a security issue. The recommended option to allow for QoS/DSCP marking is by utilizing Microsoft Group Policies, called Group Policy Objects (GPOs), to allow certain applications to mark traffic based on protocol numbers and port ranges. As described earlier in this document, the identifiable traffic streams created by Cisco Jabber can be used in conjunction with GPOs to instruct the Windows operating system to mark traffic sent by a specific application (for example, CiscoJabber.exe). Like all GPOs, QoS GPOs can be configured only by an administrator, and therefore only the applications permitted by the GPOs are allowed to mark QoS via the operating system.

In most enterprises, the network administrators do not trust the QoS markings of devices that come from the data VLAN, such as PCs. Typically all traffic from the data VLANs is re-marked to a DSCP of 0 (best effort) on ingress into the access layer and then re-marked to DSCP based on other criteria such as UCP port ranges or protocol. Some enterprises with very strict OS policies and network access policies might trust the markings from operating systems over which they have full control. In this case, a QoS GPO can benefit by allowing Windows 7 or 8 operating systems to mark QoS traffic for specific applications such as a Cisco Jabber client.

For enterprises that deploy Cisco Jabber for Windows and that prefer to use GPOs to provide this level of QoS trust, this method may be an option.

Caution **Security Alert**: In a pure Windows 7 (and later versions) environment, utilizing only GPOs would allow an enterprise to unconditionally trust all data sent from those Windows devices. Because it is highly unlikely for such homogeneous environments to exist in real-world deployments, extra effort has to be taken to separate the trust model for GPO-based devices from other operating systems and devices in the same VLANs or on similar ports in the access layer.

GPOs are very similar to network access lists in how they allow the operating system to mark a specific application's QoS based on protocols, ports, and application executable. Figure 13-23 illustrates the process of QoS re-marking in Windows 7 and 8 with Jabber for Windows.

Figure 13-23 Group Policy Objects

The process illustrated in Figure 13-23 starts with a QoS Group Policy that defines the IP address range (or any), the protocol (UDP), and port ranges (audio 3000 to 3999 and video 4000 to 4999). Once configured and applied to the OS, the Jabber for Windows client downloads its configuration from Unified CM on registration and applies the SIP Profile media port range - common. From there, when a Jabber for Windows client makes a call, it utilizes the media port ranges provided from Unified CM. The GPO applied to the Windows OS, however, applies its policy to take the media traffic for audio over UDP ports 3000 to 3999 and re-mark them to EF, and over UDP ports 4000 to 4999 and re-mark them to AF41. As the traffic leaves the OS, the packets will contain the applied markings. It will be up to the network to trust these markings and allow them to progress through the network. Figure 13-23 also illustrates a similar GPO when using non-contiguous port ranges in the SIP Profile for media port range - separated ports.

Classification in Mac OS

Cisco Jabber for Mac natively requests DSCP QoS marking to the operating system, which then marks traffic without the need to configure any specific policies.

Classification in Apple iOS (iPhone and iPad)

Cisco Jabber for iPad and iPhone natively requests DSCP QoS marking to the operating system, which then marks traffic without the need to configure any specific policies.

Classification in Android

Cisco Jabber for Android natively requests DSCP QoS marking to the operating system, which then marks traffic without the need to configure any specific policies.

Endpoint Identification and Classification Considerations and Recommendations

Design and deployment considerations and recommendations:

- Use DSCP markings whenever possible because they apply to the IP layer end-to-end and are more granular and more extensible than Layer 2 markings.
- Mark as close to the endpoint as possible, preferably at the LAN switch level.
- When deploying Jabber for voice and video in an environment where SCCP-based audio endpoints are deployed, change the media port range of the Cisco Jabber endpoints to use a range outside of 16384 to 32767 (which is a hard-coded range for SCCP devices). This is to avoid any potential overlap when creating network policies to re-mark DSCP based on the UDP port range. For example, use ports 3000 to 3999 for voice-only (video disabled) Jabber clients and 3000 to 4999 for video-enabled Cisco Jabber endpoints.
- When trying to minimize the number of media ports used by the Cisco Jabber client, use a minimum range of 100 ports. This is to ensure that there are enough ports for all of the streams, such as RTCP, RTP for audio and video, BFCP, and RTP for secondary video for desktop sharing sessions, as well as to avoid any overlap with other applications on the same computer.
- When deploying Enhanced Locations CAC, over-provision the audio class (EF) to account for the audio of video from Jabber clients that will be marked EF and *not* AF41.

Deploying QoS for Cisco Jabber clients can be achieved by mapping identifiable media and signaling streams with Layer 4 port ranges to Layer 3 DSCP values. Mapping identifiable media and signaling streams can be done for any Jabber client by using network access control lists (ACLs) or by using the operating system and then allowing the PC, Mac, or hand-held device's QoS markings to pass through the network by trusting the QoS markings. Combining both methods is not advisable because the network ACL method will simply override the OS trust method and force the re-marking of all audio, thus rendering useless the goal of using the trust method.

WAN Queuing and Scheduling

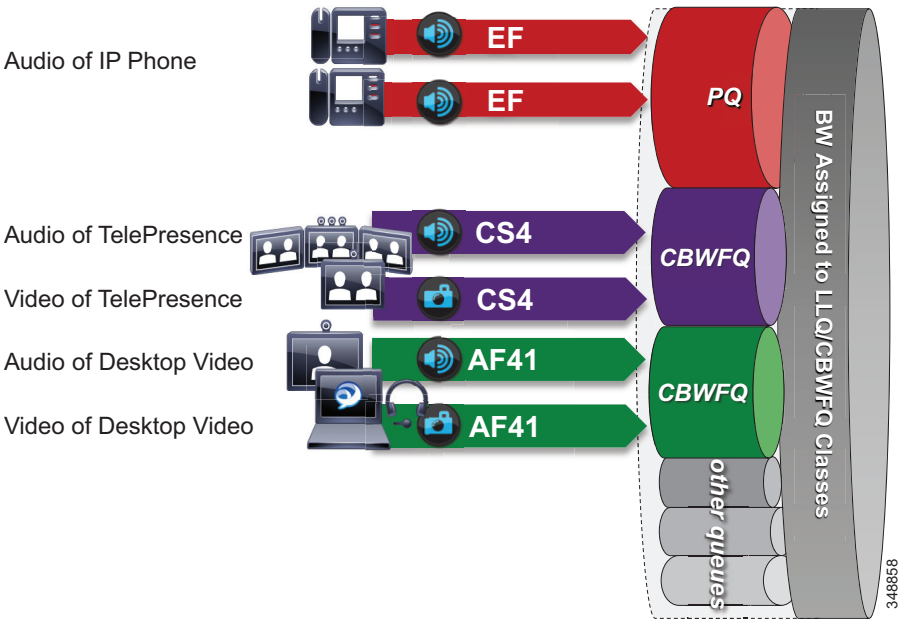
Cisco's recommendations on QoS have evolved slightly over the past few years with regards to video. Historically two types of video have been classified: desktop video and immersive TelePresence video. As discussed in the section on Identification and Classification, page 13-17, Unified CM has the ability to differentiate the video endpoint types and the video streams from these endpoints. This provides the network administrator the ability to treat the video from these two types of endpoints differently. Historically the recommended DSCP markings have been AF41 for desktop video and CS4 for TelePresence video (immersive video). These values were in line with RFC 4594. Figure 13-24 illustrates a typical approach to classification and scheduling in the WAN. This identification and classification approach has been employed for a number of years but has some shortcomings when these two classes of traffic are applied to separate rate-based queues such as Class Based Weighted Fair Queues in Cisco IOS.

Note This section discusses different Cisco IOS queuing and scheduling technologies that are covered in more detail in the section on WAN Quality of Service (QoS), page 3-37. This section discusses some of these technologies with the assumption that they are well understood technologies, and the discussion herein focuses on the best practices and recommendations for using these various Cisco IOS queuing and scheduling mechanisms.

Dual Video Queue Approach

In this approach to scheduling and queuing the traffic in the WAN, audio of a voice call is marked as EF and placed into a Priority Queue (PQ) with a strict policer on how much bandwidth the PQ can allocate to this traffic. Video calls are separated into two classes, an AF41 class for desktop type video and a CS4 class for TelePresence video (immersive). Each of these classes is put into a separate Class Based Weighted Fair Queue.

Figure 13-24 Dual Video Queue Approach



The dual video queue approach has the following shortcomings:

- Different queues for TelePresence (immersive) video and desktop video
- Complex provisioning — Requires managing multiple video queues and separating bandwidth allocations for each type of video rather than for video as a whole
- Sub-optimal bandwidth usage — When video for one class is not using all of its bandwidth, the remainder of the bandwidth becomes available to all of the other queues on the interface and not just the other video queue. Thus, it is not optimal for two different classes of video to share the total video bandwidth allocation effectively.

Other considerations of this approach with regard to the audio portion of a video call:

- Audio of a video call can be impacted by packet loss in the video queue.
  - Same DSCP for audio and video streams of a video call

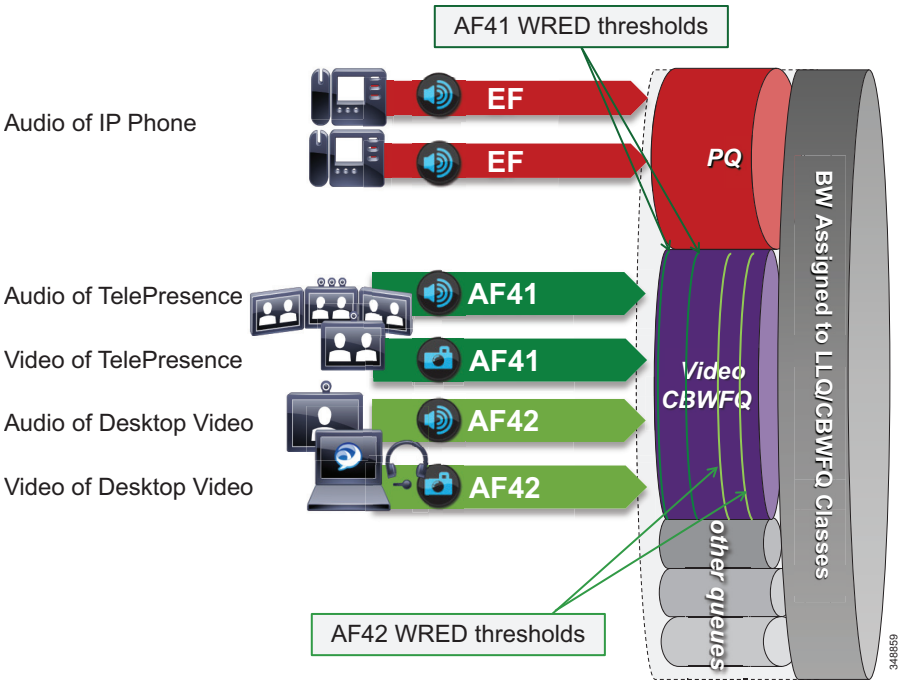
    By default both audio and video of a video call are marked with the same DSCP value. As a result, both audio and video streams are equally impacted during congestion of the video queue. When video experiences packet loss, some video quality degradation can take place during the time that it takes for the video endpoints to rate adapt down to an acceptable level until packet loss is no longer experienced. Audio is a constant bit rate medium and does not have the same abilities for rate adaptation as video does. Thus, for audio this degradation can mean that the users are no longer able to communicate until the packet loss in the video queue is under control. Impacting audio has a greater effect on user experience than does impacting video. When video is impacted, users can still carry on a meeting or conversation while video is experiencing packet loss. See the section on Audio versus Video, page 13-6, for more information on the characteristics of both media.
  - Audio and video streams of a video call were traditionally marked with the same DSCP value in order to ensure that there was not a large delay variance between the two streams, otherwise video endpoints would not be able to sync audio and video correctly. With the implementation of RTCP in all Cisco endpoints, this is no longer a concern because RTCP can ensure the proper sync between audio and video of a video call. Of course, this requires RTCP to be enabled on the video endpoints.
- Audio stream classification for untrusted devices cannot be distinguished between voice-only calls and video calls.
  - Media stream identification is difficult for untrusted endpoints and clients. As discussed earlier, when the endpoint or client is not trusted, alternative methods for identification are required. With alternative methods such as access lists, it is difficult if not impossible in most cases to differentiate the audio of a voice-only call from the audio of a video call to classify those two types of audio differently. Therefore, all audio from both types of calls would have to be marked with a single DSCP value. This makes creating a holistic approach to uniform marking more difficult.

Single Video Queue Approach

A newer recommendation for managing multiple types of video across an integrated collaboration media and data network is to use a single rate-based queue with multiple DSCPs with differing drop probabilities. In this new approach to scheduling video traffic in the WAN, the single video queue is configured with 2 or 3 AF4 drop probabilities using AF41, AF42, and AF43 – where AF43 has a higher drop precedence or probability than AF42, and AF42 has a higher drop precedence or probability than AF41. The premise behind a single video queue with this service class with hierarchical drop precedence is that, when one class of video is not using the bandwidth within the queue, the rest of the queue bandwidth is available for the other DSCP. This solves one of the major shortcoming of sub-optimal bandwidth utilization of the previous queuing approach with CS4 TelePresence video and AF41 desktop video in two separate rate-based queues.

Many different strategies for optimized video bandwidth utilization can be designed based on this single video queue with hierarchical DSCP drop probabilities. A simple example of this new QoS queuing approach can be illustrated by using the same two types of video, TelePresence video and desktop video, with two DSCP values of AF41 and AF42 in a single Class Based Weighted Fair Queue (CBWFQ). Figure 13-25 illustrates this approach.

Figure 13-25 Single Video Queue Approach



In Figure 13-25 the audio of a voice call is marked as EF and placed into a Priority Queue (PQ) with a strict policer on how much bandwidth the PQ can allocate to this traffic. Video calls are separated into two classes, AF41 for TelePresence video and AF42 for Desktop video. Using a CBWFQ with Weighted Random Early Detection (WRED), the administrator can adjust the drop precedence of AF42 over AF41, thus ensuring that during times of congestion when the queue is filling up, AF42 packets are dropped from the queue at a higher probability than AF41. See the section on WAN Quality of Service (QoS), page 3-37, for more detail on the function of WRED.

This example illustrates how an administrator using a single CBWFQ with DSCP-based WRED for all video can protect one type of video (TelePresence video) from packet loss over another type of video (Desktop) during periods of congestion. With this "single video queue approach," unlike the dual video queue approach, when one type of video is not using bandwidth in the queue, the other type of video gains full access to the entire queue bandwidth if and when needed. This is a significant point when looking to deploy pervasive video.

Considerations for Audio of Video Calls

The above single video queue example simply illustrates a point about how unused bandwidth from one class of video can be used fully by another class of video if both classes are in the same CBFWQ. This solves one of the shortcomings of the dual video queue approach. However, this does not address the other considerations for the audio portion of a video call, which as mentioned, has two main shortcomings:

- Audio of a video call can be impacted by packet loss in the video queue.
- Audio stream classification for untrusted devices cannot be distinguished between voice-only calls and video calls.

A strategy to address these deficiencies is to ensure that all audio is marked with a single value of Expedited Forwarding (EF) across the solution. In this way, whether the audio stream is associated to a voice-only call or a video call, it is always marked to the same single value. In this way, audio of a video call will be prioritized above the video and not subject to any packet loss in the video queue. It also solves the identification issue with untrusted devices such as Jabber clients. Because the marking of the client is not trusted by the network access layer, there is no effective way of distinguishing the audio stream of a voice-only call from the audio of a video call in the network. Thus, moving to this new model where all audio is marked with the same single value simplifies the network prioritization and treatment of the traffic.

Note See the section on Trusted Endpoints, page 13-19, for information on how trusted endpoints acquire DSCP and how to set the DSCP for the audio portion of a video or TelePresence endpoint, and for information on which endpoints support this differentiation. Also, the section on Untrusted Endpoints and Clients, page 13-23, shows how to set DSCP for Jabber clients.

Achieving this holistically across the entire solution depends on a number of conditions that are required to achieve marking all audio to a DSCP of EF:

- The endpoint must support the **DSCP for Audio Portion of Video/TelePresence Call** QoS setting in Unified CM to be able to mark all audio as EF. See Table 13-8 for details on endpoint support.
- Jabber clients can support marking all audio as EF in a trusted or untrusted implementation.
- Enhanced Locations CAC can be implemented in conjunction with marking all audio as EF. ELCAC relies on the correct DSCP setting to ensure protection of the queues that voice and video CAC pools represent. Changing the DSCP of audio streams of the video calls requires updating how ELCAC deducts bandwidth for video calls. This can be done by setting the service parameter under the Call Admission Control section of the CallManager service, called **Deduct Audio Bandwidth from Audio Pool for Video Call**. This parameter can be set to true or false:
  - **True**: Cisco Unified CM splits the audio and video bandwidth allocations for video calls into separate pools. The bandwidth allocation for the audio portion of a video call is deducted from the audio pool, while the video portion of a video call is deducted from the video pool.
  - **False**: Cisco Unified CM applies the legacy behavior, which is to deduct the audio and video bandwidth allocations of a video call from the video pool. This is the default setting.

For more information on the admission control aspects of marking all audio of video to EF, see the ELCAC section on Deducting all Audio from the Voice Pool, page 13-50.

Opportunistic Video

When attempting to deploy video pervasively across the organization, bandwidth constraints typically determine the level of video resolution can be achieved during the busiest hour of the day based on the bandwidth available and the number of video calls during that busy hour. To address this challenge, a type of video can be targeted as opportunistic video using a single video queue with DSCP-based WRED coupled with a strategy for identification and classification of collaboration media.

Opportunistic video means achieving the best video quality based on the WAN bandwidth resources available at any given time. To achieve this, a number of requirements must be met:

- Selecting a group of video endpoints to be opportunistic
- Ensuring the WAN is configured with a single video queue using DSCP-based WRED with AF4 DSCP class servicing with drop precedence of AF41, AF42, and AF43 (only two DSCPs are required)
- Identifying and classifying the video of opportunistic endpoints with AF42
- Identifying and classifying all other video endpoints with AF41

Provisioning and Admission Control

Provisioning bandwidth and ensuring the correct bit rate is negotiated between various groups of endpoints, are important aspects of bandwidth management. In a Unified CM environment, bit rate is negotiated via Unified CM, which uses a concept of regions to set the maximum audio and maximum video bit rates for any given call flow. This section focuses on the maximum bit rate for video and TelePresence.

Unified CM Regions

Unified CM locations (see Enhanced Locations Call Admission Control, page 13-39) work in conjunction with *regions* to define the characteristics of a call flow. Regions define the type of compression or bit rate (8 kbps or G.729, 64 kbps or G.722/G.711, and so forth) that is used between any two devices. Location links define the amount of available bandwidth for the path between devices. Each device and trunk in the system is assigned to both a region (by means of a device pool) and a location (by means of a device pool or by direct configuration on the device itself):

- Regions allow the bandwidth of video calls to be set. The audio limit on the region can result in filtering out codecs with higher bit rates. However, for video calls, the video limit constrains the quality (resolution and transmission rate) of the video.
- Locations define the amount of total bandwidth available for all calls on that link. When a call is made on a link, the regional value for that call must be subtracted from the total bandwidth allowed for that link.

Building a region matrix to manage maximum video bit rate (video resolution) for groups of devices can assist in ensuring that certain groups of devices do not over-saturate the network bandwidth. Some guidelines for creating a region matrix include:

- Group devices into maximum video bit rate categories.
- The smaller the number of groups, the easier it is to calculate bandwidth requirements.
- Consider the default region settings to simplify the matrix and provide intra-region and inter-region defaults.

For more about region settings, see the section on Enhanced Locations Call Admission Control, page 13-39.

Table 13-9 shows an example of a maximum video bit rate region matrix for four groups of devices.

Note Table 13-9 is only an example of how to group devices and what maximum bit rate might be suggested for a general resolution between the groups of devices.

Table 13-9 Example of Group Region Matrix

Endpoint Groupings	Legacy (Small Screen)	Jabber	Room System + Smart Desktop	Immersive + MCU
Legacy (Small Screen)	800 kbps	800 kbps	800 kbps	800 kbps
Jabber	800 kbps	1,500 kbps	1,500 kbps	1,500 kbps
Room System + Smart Desktop	800 kbps	1,500 kbps	2,500 kbps	2,500 kbps
Immersive + MCU	800 kbps	1,500 kbps	2,500 kbps	12,000 kbps

In Table 13-9 the four groups are:

- Legacy (Small Screen) — These could be legacy endpoints with smaller low-resolution screens or other devices to be capped at 800 kbps bit rate.
- Jabber — These would typically make up the largest group of deployed video-capable endpoints and thus benefit from the opportunistic video approach. When classified as opportunistic video, they can go up to a maximum of 1,500 kbps (720p@30fps) and would rate adapt downward based on packet loss.
- Room System + Smart Desktop — These would be room systems such as Cisco MX, SX, C, or Profile Series. Also, smart desktop endpoints would be Cisco DX and EX Series. At 2,500 kbps maximum video bit rate, these endpoints would typically be capable of 720p@30fps
- Immersive + MCU — These would be the larger Cisco TX or IX Series endpoints as well as MCUs set to a maximum of 12 Mbps, which roughly translates to 1080p@30fps with other TelePresence devices and MCUs.

Other region considerations for bandwidth provisioning:

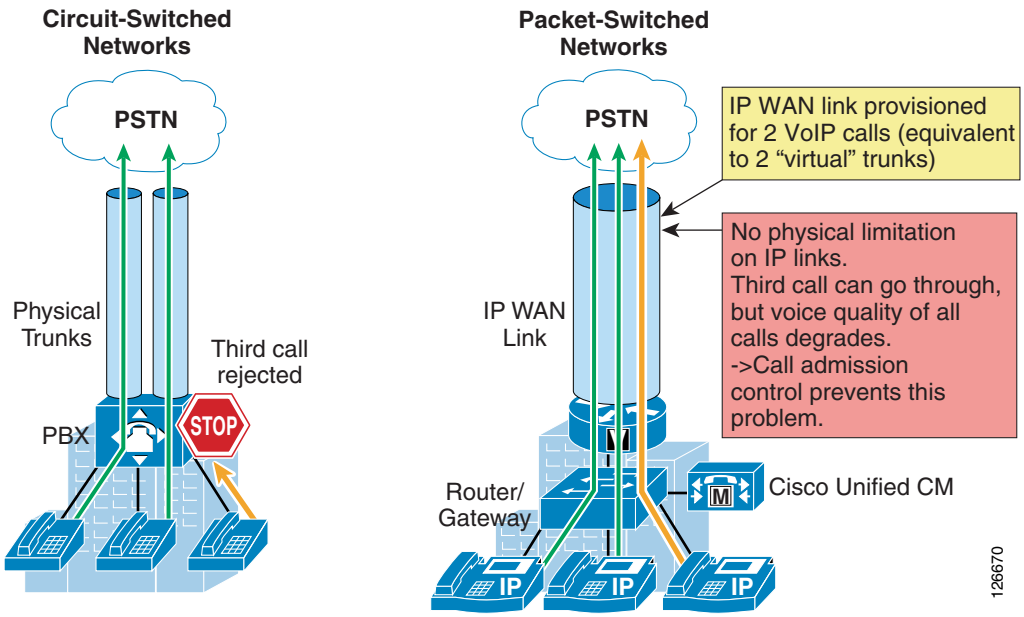
- The first consideration is whether to have different intra-region settings versus inter-region settings. This will determine whether per-site regions are required or not. The concept here is that if intra- and inter-regional audio or video bit rates are to be different, then per-site regions will be required. This augments the configuration of regions to the number of sites (N) multiplied by the number of video groups (X): N∗X = number of regions required on average. If intra- and inter-region audio and video bit rates will be the same, then only the regions for the video groups are required (X).
- Reuse regions configured for audio-only IP phones when possible.
  - Audio codec configuration is shared, so if video calls need to use different audio codecs, you need to configure new regions. For example, if voice-only devices use the G.729 audio codec over the WAN and G.711 or G.722 on the LAN, while video devices always use G7.11 or G.722, then the voice-only and video endpoints cannot share a region. Thus, each site would require a region per group of devices. Sites = N, and video region groups = 4 + voice-only region group; then N*4 is the number of regions required. Use the Prime Collaboration Provisioning tool or the Bulk Administration tool as configuration aids.
  - Per-site regions might not be needed if a single audio codec is used for both intra-region and inter-region calls as well as voice-only calls. If both audio and video endpoints use G.711 or G.722 over the WAN and LAN for voice-only or video calls, then voice-only IP phones and video endpoints could use the same region.

- Consider the default region settings to simplify the matrix. The following example illustrates possible default settings based on the region groupings in Figure 13-25. If it is desired to have the intra-region bit rate be larger than the inter-region bit rate, then per-site regions are required.
  - Default Intraregion Max Video Call Bit Rate (Includes Audio): Set to 768, sets the maximum video bit rate capability of devices for calls within a region to 768 kbps.
  - Default Interregion Max Video Call Bit Rate (Includes Audio): Set to 768, sets the maximum video bit rate capability of devices for calls between regions to 768 kbps.
  - Default Intraregion Max Immersive Video Call Bit Rate (Includes Audio): Set to 12000, sets the maximum video bit rate capability of devices for calls within a region to 12,000 kbps.
  - Default Interregion Max Immersive Video Call Bit Rate (Includes Audio): Set to 12000, sets the maximum video bit rate capability of devices for calls between regions to 12,000 kbps.
  - In addition to the defaults, 4 regions should be set up, one for each group of video endpoints.

Enhanced Locations Call Admission Control

The call admission control function can be an important component of a Collaboration system, especially when multiple sites are connected through an IP WAN and limited bandwidth resources are available for audio and video calls. To better understand what call admission control does and why it is needed, consider the example in Figure 13-26.

Figure 13-26 Why Call Admission Control is Needed



As shown on the left side of Figure 13-26, traditional TDM-based PBXs operate within circuit-switched networks, where a circuit is established each time a call is set up. As a consequence, when a legacy PBX is connected to the PSTN or to another PBX, a certain number of physical trunks must be provisioned. When calls have to be set up to the PSTN or to another PBX, the PBX selects a trunk from those that are available. If no trunks are available, the call is rejected by the PBX and the caller hears a network-busy signal.

Now consider the IP connected Unified Communications system shown on the right side of Figure 13-26. Because it is based on a packet-switched network (the IP network), no circuits are established to set up an IP telephony call. Instead, the IP packets containing the voice samples are simply routed across the IP network together with other types of data packets. Quality of Service (QoS) is used to differentiate the voice packets from the data packets, but bandwidth resources, especially on IP WAN links, are not infinite. Therefore, network administrators dedicate a certain amount of "priority" bandwidth to voice traffic on each IP WAN link. However, once the provisioned bandwidth has been fully utilized, the IP telephony system must reject subsequent calls to avoid oversubscription of the priority queue on the IP WAN link, which would cause quality degradation for all voice calls. This function is known as call admission control, and it is essential to guarantee good voice and video quality in a multisite deployment involving an IP WAN.

To preserve a satisfactory end-user experience, the call admission control function should always be performed during the call setup phase so that, if network resources are not available, a message can be presented to the end-user or the call can be rerouted across a different network (such as the PSTN).

This chapter discusses the following main topics:

- Call Admission Control Architecture, page 13-40

  This section describes the call admission control mechanism available through Cisco Unified Communications Manager called Enhanced Location Call Admission Control. For information regarding Cisco IOS gatekeeper, RSVP, and RSVP SIP Preconditions, refer to the *Call Admission Control* chapter of the *Cisco Unified Communications System 9.0 SRND*, available at

  https://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/9x/cac.html

- Design Considerations for Call Admission Control, page 13-73

  This section shows how to apply Enhanced Location Call Admission Control based on the IP WAN topology.

Call Admission Control Architecture

This section provides design and configuration guidelines for Enhanced Location Call Admission Control based on Cisco Unified CM.

Unified CM Enhanced Location Call Admission Control

Cisco Unified CM provides Enhanced Location Call Admission Control (ELCAC) to support complex WAN topologies as well as distributed deployments of Unified CM for call admission control where multiple clusters manage devices in the same physical sites using the same WAN uplinks. The Enhanced Location CAC feature also supports immersive video, allowing the administrator to control call admissions for immersive video calls such as TelePresence separately from other video calls.

To support more complex WAN topologies Unified CM has implemented a location-based network modeling functionality. This provides Unified CM with the ability to support multi-hop WAN connections between calling and called parties. This network modeling functionality has also been incrementally enhanced to support multi-cluster distributed Unified CM deployments. This allows the administrator to effectively "share" locations between clusters by enabling the clusters to communicate with one another to reserve, release, and adjust allocated bandwidth for the same locations across clusters. In addition, an administrator has the ability to provision bandwidth separately for immersive video calls such as TelePresence by allocating a new field to the Location configuration called **immersive video bandwidth**.

There are also tools to administer and troubleshoot Enhanced Location CAC. The CAC enhancements and design are discussed in detail in this chapter, but the troubleshooting and serviceability tools are discussed in separate product documentation.

Network Modeling with Locations, Links, and Weights

Enhanced Location CAC is a model-based static CAC mechanism. Enhanced Location CAC involves using the administration interface in Unified CM to configure Locations and Links to model the "Routed WAN Network" in an attempt to represent how the WAN network topology routes media between groups of endpoints for end-to-end audio, video, and immersive calls. Although Unified CM provides configuration and serviceability interfaces in order to model the network, it is still a "static" CAC mechanism that does not take into account network failures and network protocol rerouting. Therefore, the model needs to be updated when the WAN network topology changes. Enhanced Location CAC is also call oriented, and bandwidth deductions are per-call not per-stream, so asymmetric media flows where the bit-rate is higher in one direction than in the other will always deduct for the highest bit rate. In addition, unidirectional media flows will be deducted as if they were bidirectional media flows.

Enhanced Location CAC incorporates the following configuration components to allow the administrator to build the network model using Locations and Links:

- Locations — A Location represents a LAN. It could contain endpoints or simply serve as a transit location between links for WAN network modeling. For example, an MPLS provider could be represented by a Location.
- Links — Links interconnect locations and are used to define bandwidth available between locations. Links logically represent the WAN link and are configured in the Location user interface (UI).
- Weights — A weight provides the relative priority of a link in forming the *effective path* between any pair of locations. The effective path is the path used by Unified CM for the bandwidth calculations, and it has the least cumulative weight of all possible paths. Weights are used on links to provide a "cost" for the "effective path" and are pertinent only when there is more than one path between any two locations.
- Path — A path is a sequence of links and intermediate locations connecting a pair of locations. Unified CM calculates least-cost paths (lowest cumulative weight) from each location to all other locations and builds a map of the various paths. Only one "effective path" is used between any pair of locations.
- Effective Path — The effective path is the path with the least cumulative weight.
- Bandwidth Allocation — The amount of bandwidth allocated in the model for each type of traffic: audio, video, and immersive video (TelePresence).
- Location Bandwidth Manager (LBM) — The active service in Unified CM that assembles a network model from configured location and link data in one or more clusters, determines the effective paths between pairs of locations, determines whether to admit calls between a pair of locations based on the availability of bandwidth for each type of call, and deducts (reserves) bandwidth for the duration of each call that is admitted.
- Location Bandwidth Manager Hub — A Location Bandwidth Manager (LBM) service that has been designated to participate directly in intercluster replication of fixed locations, links data, and dynamic bandwidth allocation data. LBMs assigned to an LBM hub group discover each other

through their common connections and form a fully-meshed intercluster replication network. Other LBM services in a cluster with an LBM hub participate indirectly in intercluster replication through the LBM hubs in their cluster.

Locations and Links

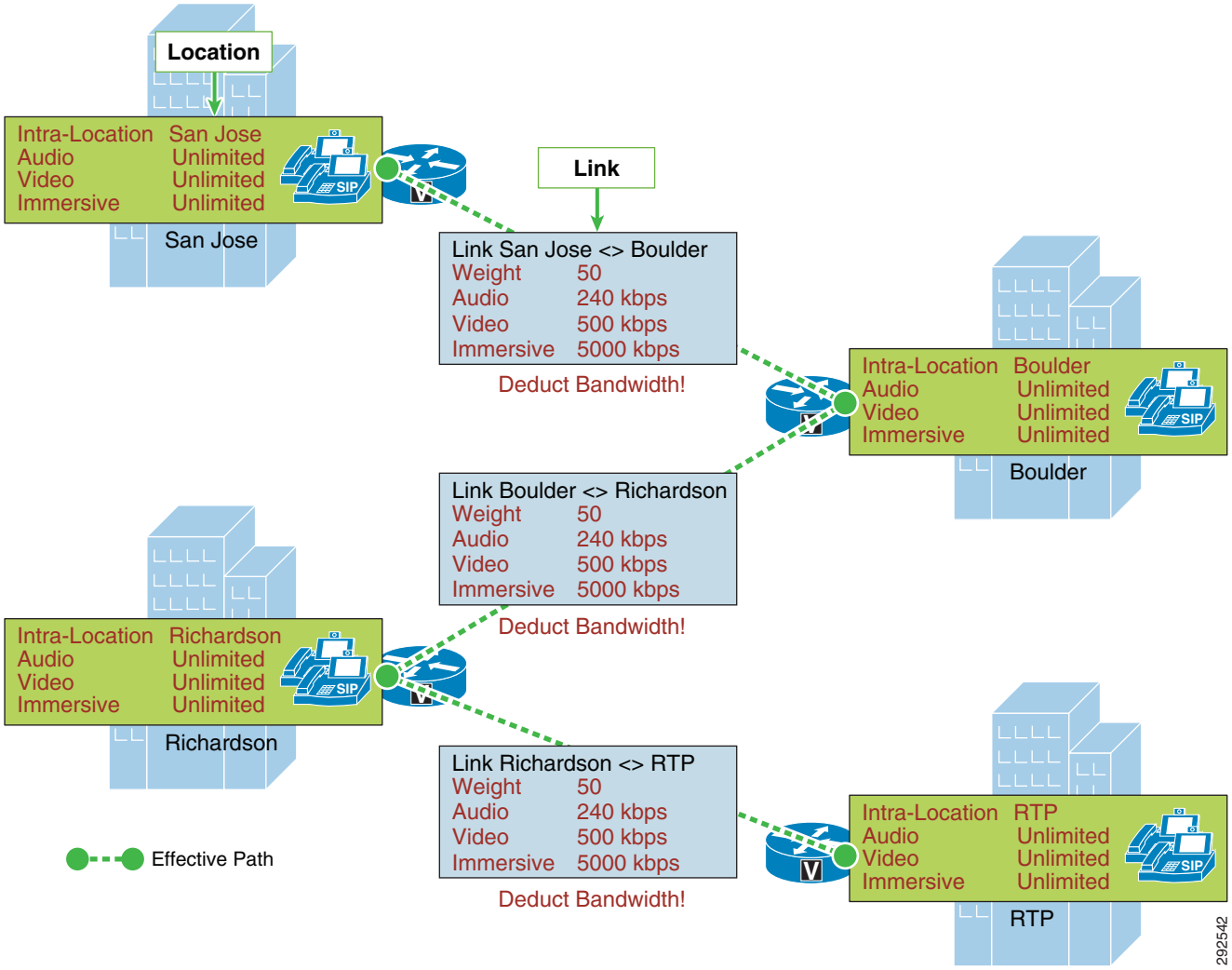
Unified CM uses the concept of locations to represent a physical site and to create an association with media devices such as endpoints, voice messaging ports, trunks, gateways, and so forth, through direct configuration on the device itself, through a device pool, or even through device mobility. Unified CM also uses a new location configuration parameter called *links*. Links interconnect locations and are used to define bandwidth available between locations. Links logically represent the WAN links. This section describes locations and links and how they are used.

The location configuration itself consists of three main parts: links, intra-location bandwidth parameters, and RSVP locations settings. The RSVP locations settings are not considered here for Enhanced Location CAC because they apply only to RSVP implementations. In the configuration, the link bandwidth parameters are displayed first while the intra-location bandwidth parameters are hidden and displayed by selecting the **Show advanced** link.

The intra-location bandwidth parameters allow the administrator to configure bandwidth allocations for three call types: audio, video, and immersive. They limit the amount of traffic within, as well as to or from, any given location. When any device makes or receives a call, bandwidth is deducted from the applicable bandwidth allocation for that call type. This feature allows administrators to limit the amount of bandwidth used on the LAN or transit location. In most networks today that consist of Gigabit LANs, there is little or no reason to limit bandwidth on those LANs.

The link bandwidth parameters allow the administrator to characterize the provisioned bandwidth for audio, video, and immersive calls between "adjacent locations" (that is, locations that have a link configured between them). This feature offers the administrator the ability to create a string of location pairings in order to model a multi-hop WAN network. To illustrate this, consider a simple three-hop WAN topology connecting four physical sites, as shown in Figure 13-27. In this topology we want to create links between San Jose and Boulder, between Boulder and Richardson, and between Richardson and RTP. Note that when we create a link from San Jose to Boulder, for example, the inverse link (Boulder to San Jose) also exists. Therefore, the administrator needs to create the link pairing only once from either location configuration page. In the example in Figure 13-27, each of the three links has the same settings: a weight of 50, 240 kbps of audio bandwidth, 500 kbps of video bandwidth, and 5000 kbps (or 5 Mbps) of immersive bandwidth.

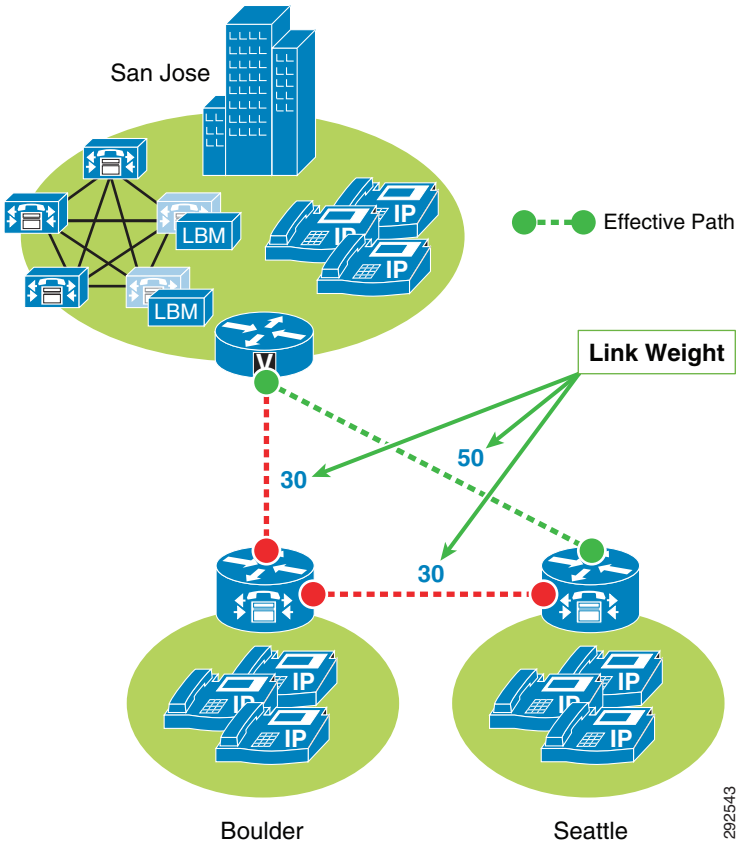
Figure 13-27 Simple Link Example with Three WAN Hops



When a call is made between San Jose and RTP, Unified CM calculates the bandwidth of the requested call, which is determined by the region pair between the two devices (see Locations, Links, and Region Settings, page 13-46) and verifies the effective path between the two locations. That is to say, Unified CM verifies the locations and links that make up the path between the two locations and accordingly deducts bandwidth from each link and (if applicable) from each location in the path. The intra-location bandwidth also is deducted along the path if any of the locations has configured a bandwidth value other than unlimited.

Weight is configurable on the link only and provides the ability to force a specific path choice when multiple paths between two locations are available. When multiple paths are configured, only one will be selected based on the cumulative weight, and this path is referred to as the *effective path*. This weight is static and the effective path does not change dynamically. Figure 13-28 illustrates weight configured on links between three locations: San Jose, Boulder, and Seattle.

Figure 13-28 Cumulative Path Weights



San Jose to Seattle has two paths, one direct link between the locations and another path through the Boulder location (link San Jose/Boulder and link Boulder/Seattle). The weight configured on the direct link between San Jose and Seattle is 50 and is less than the cumulative weight of links San Jose/Boulder and Boulder/Seattle which is 60 (30+30). Thus, the direct link is chosen as the effective path because the cumulative link weight is 50.

When you configure a device in Unified CM, the device can be assigned to a location. A location can be configured with links to other locations in order to build a topology. The locations configured in Unified CM are virtual locations and not real, physical locations. As mentioned, Unified CM has no knowledge of the actual physical topology of the network. Therefore, any changes to the physical network must be made manually in Unified CM to map the real underlying network topology with the Unified CM locations model. If a device is moved from one physical location to another, the system administrator must either perform a manual update on its location configuration or else implement the device mobility feature so that Unified CM can correctly calculate bandwidth allocations for calls to and from that device. Each device is in location **Hub_None** by default. Location Hub_None is an example location that typically serves as a hub linking two or more locations, and it is configured by default with unlimited intra-location bandwidth allocations for audio, video, and immersive bandwidth.

Unified CM allows you to define separate voice, video, and immersive video bandwidth pools for each location and link between locations. Typically the locations intra-location bandwidth configuration is left as a default of **Unlimited** while the link between locations is set to a finite number of kilobits per second (kbps) to match the capacity of a WAN links between physical sites. If the location's intra-location audio, video, and immersive bandwidths are configured as **Unlimited**, there will be unlimited bandwidth available for all calls (audio, video, and immersive) within that location and

transiting that location. On the other hand, if the bandwidth values are set to a finite number of kilobits per second (kbps), Unified CM will track all calls within the location and all calls that use the location as a transit location (a location that is in the calculation path but is not the originating or terminating location in the path).

For video calls, the video location bandwidth takes into account both the audio and the video portions of the video call. Therefore, for a video call, no bandwidth is deducted from the audio bandwidth pool. The same applies to immersive video calls.

The devices that can specify membership in a location include:

- IP phones
- CTI ports
- H.323 clients
- CTI route points
- Conference bridges
- Music on hold (MoH) servers
- Gateways
- Trunks
- Media termination point (via device pool)
- Trusted relay point (via device pool)
- Annunciator (via device pool)

The Enhanced Location Call Admission Control mechanism also takes into account the mid-call changes in call type. For example, if an inter-site video call is established, Unified CM will subtract the appropriate amount of video bandwidth from the respective locations and links in the path. If this video call changes to an audio-only call as the result of a transfer to a device that is not capable of video, Unified CM will return the allocated bandwidth to the video pool and allocate the appropriate amount of bandwidth from the audio pool along the same path. Calls that change from audio to video will cause the opposite change of bandwidth allocation.

Table 13-10 lists the amount of bandwidth requested by the static locations algorithm for various call speeds. For an audio call, Unified CM counts the media bit rates plus the IP and UDP overhead. For example, a G.711 audio call consumes 80 kbps (64k bit rate + 16k IP/UDP headers) deducted from the location's and link's audio bandwidth allocation. For a video call, Unified CM counts only the media bit rates for both the audio and video streams. For example, for a video call at a bit rate of 384 kbps, Unified CM will allocate 384 kbps from the video bandwidth allocation.

Table 13-10 Amount of Bandwidth Requested by the Locations and Links Bandwidth Deduction Algorithm

Call Speed	Static Location and Link Bandwidth Value
G.711 audio call (64 kbps)	80 kbps
G.729 audio call (8 kbps)	24 kbps
128 kbps video call	128 kbps
384 kbps video call	384 kbps
512 kbps video call	512 kbps
768 kbps video call	768 kbps

For a complete list of codecs and location and link bandwidth values, refer to the bandwidth calculations information in the *Call Admission Control* section of the *Cisco Unified Communications Manager System Guide*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

For example, assume that the link configuration for the location Branch 1 to Hub_None allocates 256 kbps of available audio bandwidth and 384 kbps of available video bandwidth. In this case the path from Branch 1 to Hub_None can support up to three G.711 audio calls (at 80 kbps per call) or ten G.729 audio calls (at 24 kbps per call), or any combination of both that does not exceed 256 kbps. The link between locations can also support different numbers of video calls depending on the video and audio codecs being used (for example, one video call requesting 384 kbps of bandwidth or three video calls with each requesting 128 kbps of bandwidth).

When a call is placed from one location to the other, Unified CM deducts the appropriate amount of bandwidth from the effective path of locations and links from one location to another. Using Figure 13-27 as an example, a G.729 call between San Jose and RTP locations causes Unified CM to deduct 24 kbps from the available bandwidth at the links between San Jose and Boulder, between Boulder and Richardson, and between Richardson and RTP. When the call has completed, Unified CM returns the bandwidth to those same links over the effective path. If there is not enough bandwidth at any one of the links over the path, the call is denied by Unified CM and the caller receives the network busy tone. If the calling device is an IP phone with a display, that device also displays the message "Not Enough Bandwidth."

When an inter-location call is denied by call admission control, Unified CM can automatically reroute the call to the destination through the PSTN connection by means of the Automated Alternate Routing (AAR) feature. For detailed information on the AAR feature, see Automated Alternate Routing, page 14-79.

Note AAR is invoked only when Enhanced Location Call Admission Control denies the call due to a lack of network bandwidth along the effective path. AAR is not invoked when the IP WAN is unavailable or other connectivity issues cause the called device to become unregistered with Unified CM. In such cases, the calls are redirected to the target specified in the Call Forward No Answer field of the called device.

It is also worth noting that video devices can be enabled to **Retry Video Call as Audio** if a video call between devices fails CAC. This option is configured on the video endpoint configuration page in Unified CM and is applicable to video endpoints or trunks receiving calls. It should also be noted that for some video endpoints **Retry Video Call as Audio** is enabled by default and not configurable on the endpoint.

Locations, Links, and Region Settings

Locations work in conjunction with regions to define the characteristics of a call over the effective path of locations and links. Regions define the type of compression or bit rate (8 kbps or G.729, 64 kbps or G.722/G.711, and so forth) that is used between devices, and location links define the amount of available bandwidth for the effective path between devices. You assign each device in the system to both a region (by means of a device pool) and a location (by means of a device pool or by direct configuration on the device itself).

You can configure locations in Unified CM to define:

- Physical sites (for example, a branch office) or transit sites (for example, an MPLS cloud) — A location represents a LAN. It could contain endpoints or simply serve as a transit location between links for WAN network modeling.
- Link bandwidth between adjacent locations — Links interconnect locations and are used to define bandwidth available between locations. Links logically represent the WAN link between physical sites.
  - Audio Bandwidth — The amount of bandwidth that is available in the WAN link for voice and fax calls being made from devices in the location to the configured adjacent location. Unified CM uses this bandwidth value for Enhanced Location Call Admission Control.
  - Video Bandwidth — The amount of video bandwidth that is available in the WAN link for video calls being made from devices in the location to the configured adjacent location. Unified CM uses this bandwidth value for Enhanced Location Call Admission Control.
  - Immersive Video Bandwidth — The amount of immersive bandwidth that is available in the WAN link for TelePresence calls being made from devices in the location to the configured adjacent location. Unified CM uses this bandwidth value for Enhanced Location Call Admission Control.
- Intra-location bandwidth
  - Audio Bandwidth — The amount of bandwidth that is available in the LAN for voice and fax calls being made from devices within the location. Unified CM uses this bandwidth value for Enhanced Location Call Admission Control.
  - Video Bandwidth — The amount of video bandwidth that is available in the LAN for video calls being made from devices within the location. Unified CM uses this bandwidth value for Enhanced Location Call Admission Control.
  - Immersive Video Bandwidth — The amount of immersive bandwidth that is available in the LAN for TelePresence calls being made from devices within the location. Unified CM uses this bandwidth value for Enhanced Location Call Admission Control.

You can configure regions in Unified CM to define:

- The Maximum Audio Bit Rate used for intraregion and interregion calls
- The Maximum Session Bit Rate for Video Calls (Includes Audio) used for intraregion and interregion calls
- The Maximum Session Bit Rate for Immersive Video Calls (Includes Audio) used for intraregion and interregion calls
- Audio codec preference lists

Unified CM Support for Locations and Regions

Cisco Unified Communications Manager supports 2,000 locations and 2,000 regions per cluster. To deploy up to 2,000 locations and regions, you must configure the following service parameters in the **Clusterwide Parameters > (System - Location and Region) and Clusterwide Parameters > (System - RSVP)** configuration menus:

- Default Intraregion Max Audio Bit Rate
- Default Interregion Max Audio Bit Rate
- Default Intraregion Max Video Call Bit Rate (Includes Audio)

- Default Interregion Max Video Call Bit Rate (Includes Audio)
- Default Intraregion Max Immersive Call Bit Rate (Includes Audio)
- Default Interregion Max Immersive Video Call Bit Rate (Includes Audio)
- Default Audio Codec Preference List between Regions
- Default Audio Codec Preference List within Regions

When adding regions, you should select **Use System Default** for the Maximum Audio Bit Rate and Maximum Session Bit Rate for Video Call values.

Changing these values from the default for individual regions has an impact on server initialization and publisher upgrade times. Hence, with a total of 2,000 regions and 2,000 locations, you can modify up to 200 regions to use non-default values. With a total of 1,000 or fewer regions and locations, you can modify up to 500 regions to use non-default values. Table 13-11 summarizes these limits.

Table 13-11 Number of Allowed Locations and Non-Default Regions

Number of non-default regions	Maximum number of regions	Maximum number of locations
0 to 200	2,000	2,000
200 to 500	1,000	1,000

Note The Maximum Audio Bit Rate is used by both voice calls and fax calls. If you plan to use G.729 as the interregion codec, use T.38 Fax Relay for fax calls. If you plan to use fax pass-through over the WAN, use audio preference lists to prefer G.729 for audio-only calls and G.711 for fax calls.

Location Bandwidth Manager

The Location Bandwidth Manager (LBM) is a Unified CM Feature Service managed from the serviceability web pages and is responsible for all of the Enhanced Location CAC bandwidth functions. The LBM can run on any Unified CM subscriber node or as a standalone service on a dedicated Unified CM node in the cluster. A minimum of one instance of LBM must run in each cluster to enable Enhanced Location CAC in the cluster. However, Cisco recommends running LBM on each subscriber node in the cluster that is also running the Cisco CallManager service.

The LBM performs the following functions:

- Assembles topology of locations and links
- Calculates the effective paths across the topology
- Services bandwidth requests from the Cisco CallManager service (Unified CM call control)
- Replicates the bandwidth information to other LBMs
- Provides configured and dynamic information to serviceability
- Updates Location Real-Time Monitoring Tool (RTMT) counters

The LBM Service is enabled by default when upgrading Cisco Unified CM from earlier releases that support only traditional Location CAC. For new installations, the LBM service must be activated manually.

During initialization, the LBM reads local locations information from the database, such as: locations audio, video, and immersive bandwidth values; intra-location bandwidth data; and location-to-location link audio, video, and immersive bandwidth values and weight. Using the link data, each LBM in a

cluster creates a local assembly of the paths from one location to every other location. This is referred to as the *assembled topology*. In a cluster, each LBM accesses the same data and thus creates the same local copy of the assembled topology during initialization.

At runtime, the LBM applies reservations along the computed paths in the local assembled topology of locations and links, and it replicates the reservations to other LBMs in the cluster. If intercluster Enhanced Location CAC is configured and activated, the LBM can be configured to replicate the assembled topology to other clusters (see Intercluster Enhanced Location CAC, page 13-51, for more details).

By default the Cisco CallManager service communicates with the local LBM service; however, LBM groups can be used to manage this communication. LBM groups provide an active and standby LBM in order to create redundancy for Unified CM call control. Figure 13-29 illustrates LBM redundancy.

Figure 13-29 Location Bandwidth Manager Redundancy

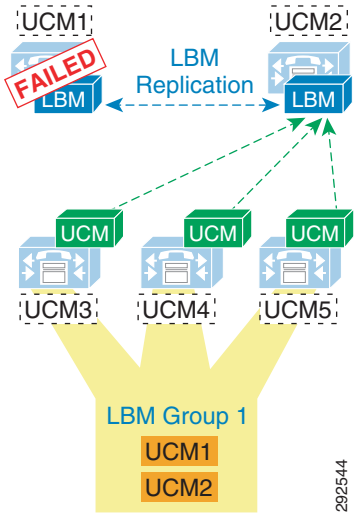


Figure 13-29 shows five Unified CM servers: UCM1 and UCM2 are dedicated LBM servers (only LBM service enabled); UCM3, UCM4, and UCM5 are Unified CM subscribers (Cisco CallManager service enabled). An LBM Group has been configured with UCM1 as active and UCM2 as standby, and it is applied to subscribers UCM3, UCM4, and UCM5. This configuration allows for UCM3, UCM4, and UCM5 to query UCM1 for all bandwidth requests. If UCM1 fails for any reason, the subscribers will fail-over to the standby UCM2. This example is used to illustrate how the LBM Group configuration functions and is not a recommended configuration (see recommendations below).

Because LBM is directly involved in processing requests for every call that is processed by the CallManager service that it is serving, it is important to follow these simple design recommendations in order to ensure optimal functionality of the LBM.

The recommended configuration is to run LBM co-resident with the Cisco CallManager service (call processing). If redundancy of the LBM service is required, it is important not to over-subscribe any given LBM. It is also important to make sure that LBM is no more than a primary and secondary in any given deployment. This means that LBM should not have the load of more than 2 CallManager services during failure scenarios, and the load of only one CallManager service during normal operation. The LBM group can be used to configure a co-resident LBM as the primary, another local (on the same LAN) LBM as secondary, and lastly the service parameter as a failsafe mechanism to ensure that all calls processed by that CallManager service do not fail. There are many reasons for these recommendations. It is difficult, at best, to determine the load of any LBM because it is directly related to the call-processing

load of the CallManager service that it is serving. There are also considerations for delay. As soon as an LBM is off-box from a CallManager service, there is an added delay caused by packetization and processing for every call serviced by the CallManager service. Compounding call-processing delay can bring the overall delay budget to an unacceptable level for any given call flow to a ringing state, and result in a poor user experience. Following these design recommendations will reduce the overall call-processing delay.

The order in which the Unified CM Cisco CallManager service uses the LBM is as follows:

- LBM Group designation
- Local LBM (co-resident)
- Service parameter **Call Treatment when no LBM available** (Default = **allow calls**)

Enhanced Location CAC Design and Deployment Recommendations and Considerations

- The Location Bandwidth Manager (LBM) is a Unified CM Feature Service. It is responsible for modeling the topology and servicing Unified CM bandwidth requests.
- LBMs within the cluster create a fully meshed communications network via XML over TCP for the replication of bandwidth change notifications between LBMs.
- Cisco recommends deploying the LBM service co-resident with a Unified CM subscriber running the Cisco CallManager call processing service.
- If redundancy is required for the LBM service, create an LBM Group for each Unified CM subscriber running the Cisco CallManager call processing service. Add the co-resident LBM service as the primary LBM, and the LBM from another Unified CM subscriber on the same LAN as a secondary LBM. This will ensure that the Cisco CallManager call processing service uses the co-resident LBM as primary, the LBM on another local (same LAN) Unified CM subscriber as secondary, and the service parameter **Call Treatment when no LBM available** as tertiary source for LBM requests.

Note Cisco recommends having LBM back up more than one Cisco CallManager service. Assuming that the LBM is handling the load of the co-resident CallManager service, and during failure of another CallManager service, this would equate to the load of 2 call processing servers on a single LBM.

- For deployments with cluster over the WAN and local failover, intracluster LBM traffic is already calculated into the WAN bandwidth calculations. See the section on clustering over the WAN Local Failover Deployment Model, page 10-47, for details on bandwidth calculations.

Deducting all Audio from the Voice Pool

Unified CM now has a feature that allows the administrator to deduct the audio bandwidth of video and TelePresence calls from the voice pool. Because ELCAC relies on the correct DSCP setting in order to ensure the protection of the queues that voice and video CAC pools represent, changing how Unified CM deducts bandwidth from the video pool requires the DSCP of audio streams of the video calls to be marked the same as the audio streams of audio-only calls. See the section on Considerations for Audio of Video Calls, page 13-35, for information about aligning admission control with QoS.

In Unified CM this feature is enabled by setting the service parameter **Deduct Audio Bandwidth from Audio Pool for Video Call** to **True** under the Call Admission Control section of the CallManager service. False is the default setting, and by default Unified CM deducts both audio and video streams of video calls from the video pool.

Intercluster Enhanced Location CAC

Intercluster Enhanced Location CAC extends the concept of network modeling across multiple clusters. In intercluster Enhanced Location CAC, each cluster manages its locally configured topology of locations and links and then propagates this local topology to other remote clusters that are part of the LBM intercluster replication network. Upon receiving a remote cluster's topology, the LBM assembles this into its own local topology and creates a global topology. Through this process the global topology is then identical across all clusters, providing each cluster a global view of enterprise network topology for end-to-end CAC. Figure 13-30 illustrates the concept of a global topology with a simplistic hub-and-spoke network topology as an example.

Figure 13-30 Example of a Global Topology for a Simple Hub-and-Spoke Network

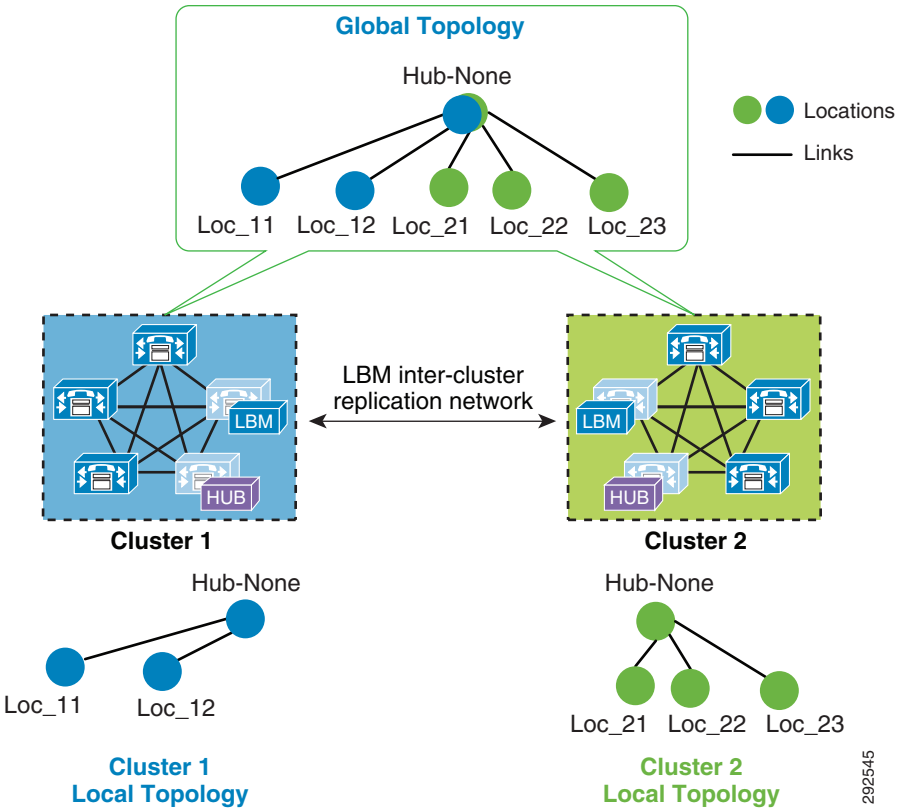


Figure 13-30 shows two clusters, Cluster 1 and Cluster 2, each with a locally configured hub-and-spoke network topology. Cluster 1 has configured Hub_None with links to Loc_11 and Loc_12, while Cluster 2 has configured Hub_None with links to Loc_21, Loc_22, and Loc_23. Upon enabling intercluster Enhanced Location CAC, Cluster 1 sends its local topology to Cluster 2, as does Cluster 2 to Cluster 1. After each cluster obtains a copy of the remote cluster's topology, each cluster overlays the remote cluster's topology over their own. The overlay is accomplished through common locations, which are locations that are configured with the same name. Because both Cluster 1 and Cluster 2 have the common location Hub_None with the same name, each cluster will overlay the other's network topology with Hub_None as a common location, thus creating a global topology where Hub_None is the hub and Loc_11, Loc_12, Loc_21, Loc_22 and Loc_23 are all spoke locations. This is an example of a simple network topology, but more complex topologies would be processed in the same way.

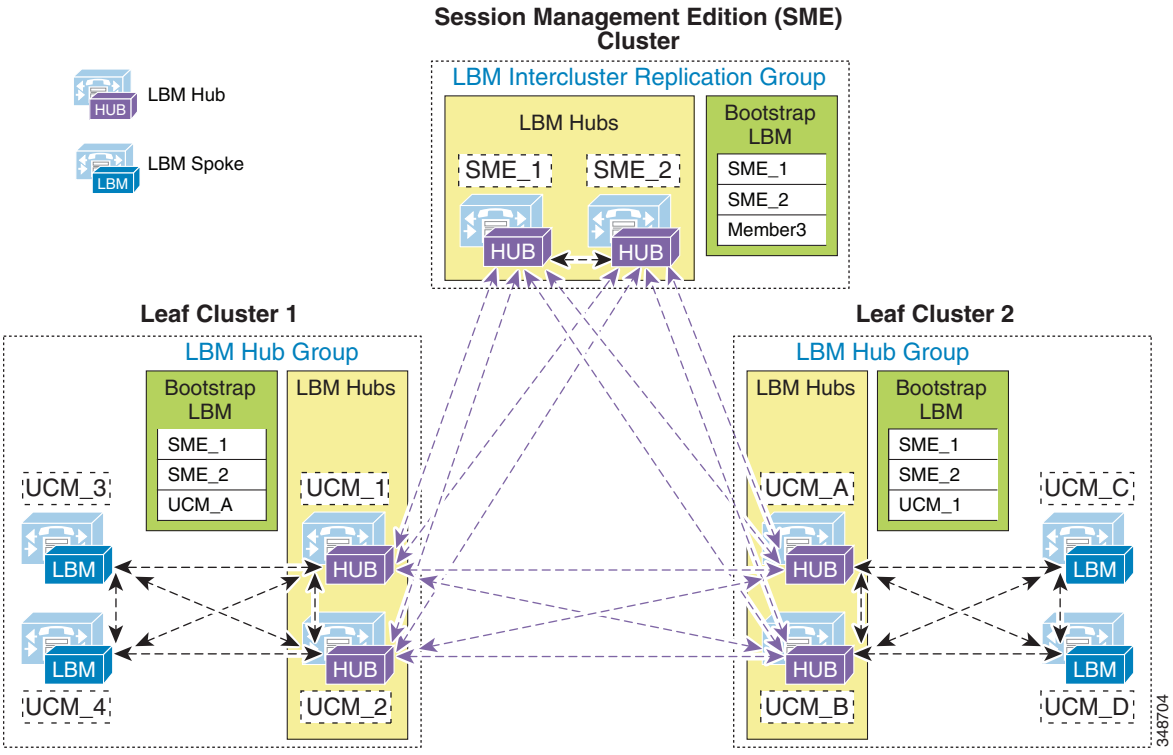
LBM Hub Replication Network

The intercluster LBM replication network is a separate replication network of designated LBMs called LBM hubs. LBM hubs create a separate full mesh with one another and replicate their local cluster's topology to other remote clusters. Each cluster effectively receives the topologies from every other remote cluster in order to create a global topology. The designated LBMs for the intercluster replication network are called *LBM hubs*. The LBMs that replicate only within a cluster are called *LBM spokes*. The LBM hubs are designated in configuration through the LBM **intercluster replication group**. The LBM role assignment for any LBM in a cluster can also be changed to a hub or spoke role in the intercluster replication group configuration (For further information on the LBM hub group configuration, refer to the Cisco Unified Communications Manager product documentation available at https://www.cisco.com/en/US/products/sw/voicesw/ps556/tsd_products_support_series_home.html.)

In the LBM intercluster replication group, there is also a concept of bootstrap LBM. Bootstrap LBMs are LBM hubs that provide all other LBM hubs with the connectivity details required to create the full-mesh hub replication network. Bootstrap LBM is a role that any LBM hub can have. If all LBM hubs point to a single LBM hub, that single LBM hub will tell all other LBM hubs how to connect to one another. Each replication group can reference up to three bootstrap LBMs.

Once the LBM hub group is configured on each cluster, the designated LBM hubs will create the full-mesh intercluster replication network. Figure 13-31 illustrates an intercluster replication network configuration with LBM hub groups set up between three clusters (Leaf Cluster 1, Leaf Cluster 2, and a Session Management Edition (SME) cluster) to form the intercluster replication network. The SME cluster is used here only as an example and is not required for this specific setup. The SME cluster could simply be another regular cluster handling endpoint registrations.

Figure 13-31 Example Intercluster Replication Network for Three Clusters



In Figure 13-31, two LBMs from each cluster have been designated as the LBM hubs for their cluster. These LBM hubs form the intercluster LBM replication network. The bootstrap LBMs configured in each LBM intercluster replication group are designated as SME_1 and SME_2. These two LBM hubs from the SME cluster serve as points of contact or bootstrap LBMs for the entire intercluster LBM replication network. This means that each LBM in each cluster connects to SME_1, replicates its local topology to SME_1, and gets the remote topology from SME_1. They also get the connectivity information for the other leaf clusters from SME_1, connect to the other remote clusters, and replicate their topologies. This creates the full-mesh replication network. If SME_1 is unavailable, the LBM hubs will connect to SME_2. If SME_2 is unavailable, Leaf Cluster 1 LBMs will connect to UCM_A and Leaf Cluster 2 LBMs will connect to UCM_1 as a backup measure in case the SME cluster is unavailable. This is just an example configuration to illustrate the components of the intercluster LBM replication network.

The LBM has the following roles with respect to the LBM intercluster replication network:

- Bootstrap LBMs
  - Remote LBM hubs responsible for interconnecting all LBM hubs in the replication network
  - Can be any hub in the network
  - Can indicate up to three bootstrap LBM hubs per LBM intercluster replication group
- LBM hubs (local LBMs)
  - Communicate directly to other remote hubs as part of the intercluster LBM replication network
- LBM spokes (local LBMs)
  - Communicate directly to local LBM hubs in the cluster and indirectly to the remote LBM hubs through the local LBM hubs
- LBM hub replication network — Bandwidth deduction and adjustment messages
  - LBM optimizes the LBM messages by choosing a sender and receiver from each cluster.
  - The LBM sender and receiver of the cluster is determined by lowest IP address.
  - The LBM hubs that receive messages from remote clusters, in turn forward the received messages to the LBM spokes in their local cluster.

LBM hubs can also be configured to encrypt their communications. This allows intercluster ELCAC to be deployed in environments where it is critical to encrypt traffic between clusters because the links between clusters might reside over unprotected networks. For further information on configuring encrypted signaling between LBM hubs, refer to the Cisco Unified Communications Manager product documentation available at

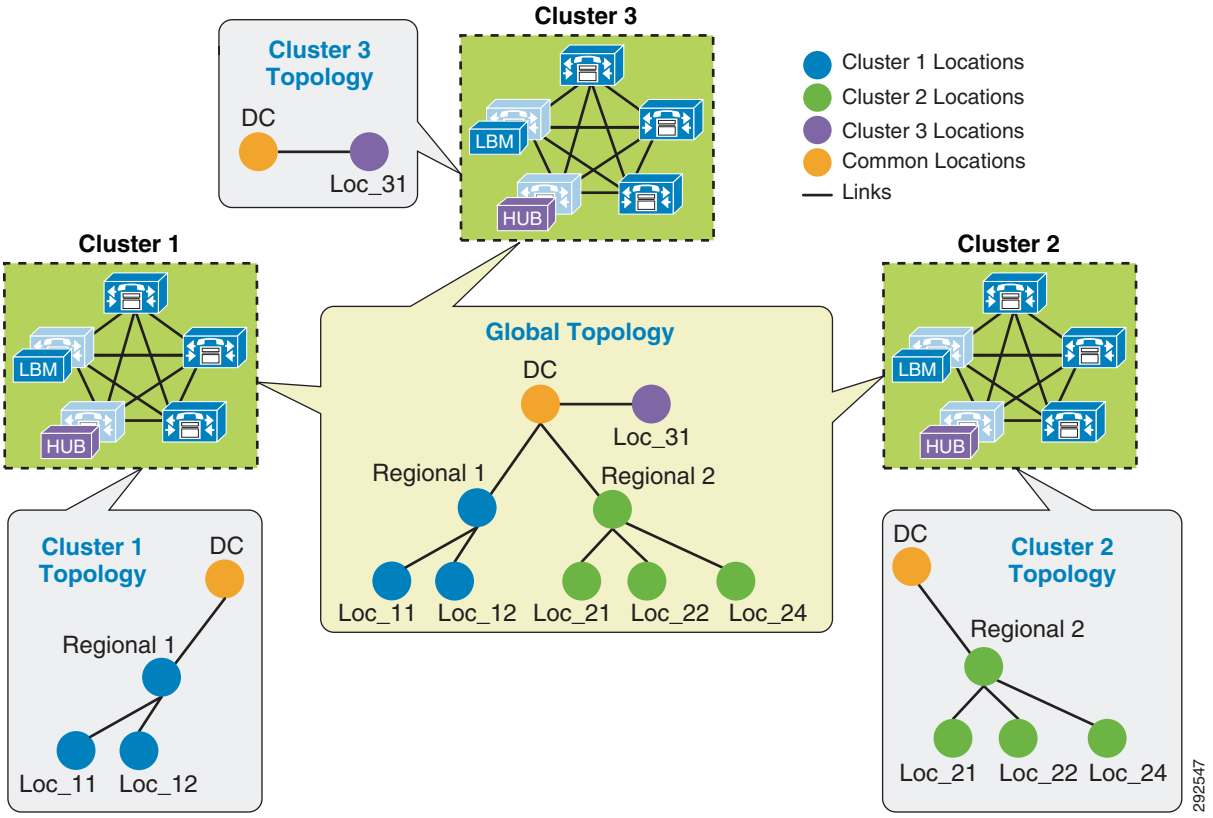
https://www.cisco.com/en/US/products/sw/voicesw/ps556/tsd_products_support_series_home.html

Common Locations (Shared Locations) and Links

As mentioned previously, common locations are locations that are named the same across clusters. Common locations play a key role in how the LBM creates the global topology and how it associates a single location across multiple clusters. A location with the same name between two or more clusters is considered the same location and is thus a shared location across those clusters. So if a location is meant to be shared between multiple clusters, it is required to have exactly the same name. After replication, the LBM will check for configuration discrepancies across locations and links. Any discrepancy in bandwidth value or weight between common locations and links can be seen in serviceability, and the LBM calculates the locations and link paths with the most restrictive values for bandwidth and the lowest value (least cost) for weight.

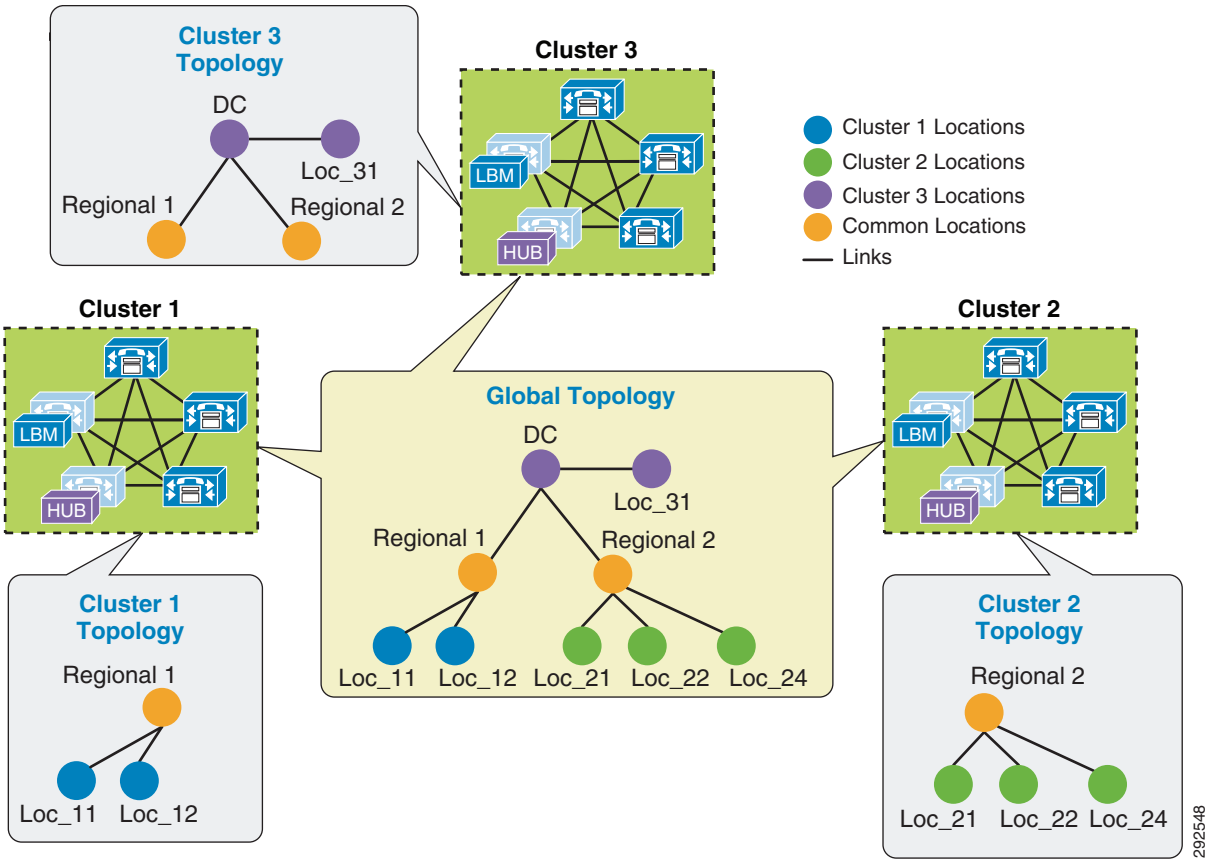
Common locations and links can be configured across clusters for a number of different reasons. You might have a number of clusters that manage devices in the same physical site and use the same WAN uplinks, and therefore the same location needs to be configured on each cluster in order to associate that location to the local devices on each cluster. You might also have clusters that manage their own topology, yet these topologies interconnect at specific locations and you will have to configure these locations as common locations across each cluster so that, when the global topology is being created, the clusters have the common interconnecting locations and links on each cluster to link each remote topology together effectively. Figure 13-32 illustrates linking topologies together and shows the common topology that each cluster shares.

Figure 13-32 Using Common Locations and Links to Create a Global Topology



In Figure 13-32, Cluster 1 has devices in locations Regional 1, Loc_11, and Loc_12, but it requires configuring DC and a link from Regional 1 to DC in order to link to the rest of the global topology. Cluster 2 is similar, with devices in Regional 2 and Loc_21, Loc_22, and Loc_23, and it requires configuring DC and a link from DC to Regional 2 to map into the global topology. Cluster 3 has devices in Loc_31 only, and it requires configuring DC and a link to DC from Loc_31 to map into Cluster 1 and Cluster 2 topologies. Alternatively, Regional 1 and Regional 2 could be the common locations configured on all clusters instead of DC, as is illustrated in Figure 13-33.

Figure 13-33 Alternative Topology Using Different Common Locations



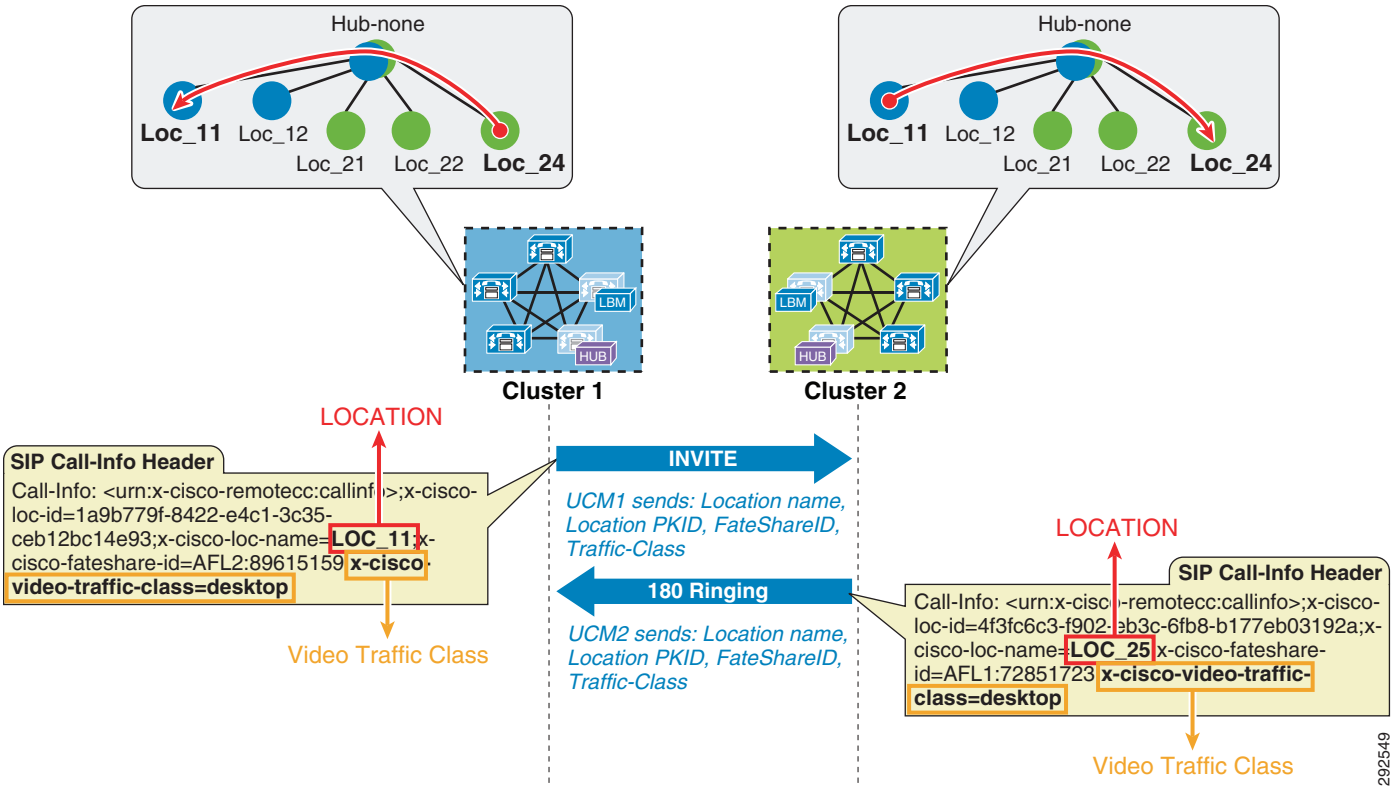
The key to topology mapping from cluster to cluster is to ensure that at least one cluster has a common location with another cluster so that the topologies interconnect accordingly.

Shadow Location

The *shadow location* is used to enable a SIP trunk to pass Enhanced Location CAC information such as location name and Video-Traffic-Class (discussed below), among other things, required for Enhanced Location CAC to function between clusters. In order to pass this location information across clusters, the SIP intercluster trunk (ICT) must be assigned to the "shadow" location. The shadow location cannot have a link to other locations, and therefore no bandwidth can be reserved between the shadow location and other locations. Any device other than a SIP ICT that is assigned to the shadow location will be treated as if it was associated to Hub_None. That is important to know because if a device other than a SIP ICT ends up in the shadow location, bandwidth deductions will be made from that device as if it were in Hub_None, and that could have varying effects depending on the location and links configuration.

When the SIP ICT is enabled for Enhanced Location CAC, it passes information in the SIP Call-Info header that allows the originating and terminating clusters to process the location bandwidth deductions end-to-end. Figure 13-34 illustrates an example of a call between two clusters and some details about the information passed. This is only to illustrate how location information is passed from cluster to cluster and how bandwidth deductions are made.

Figure 13-34 Location Information Passed Between Clusters over SIP



In Figure 13-34, Cluster 1 sends an invite to Cluster 2 and populates the call-info header with the calling parties location name and Video-Traffic-Class, among other pertinent information such as unique call-ID. When Cluster 2 receives the invite with the information, it looks up the terminating party and performs a CAC request on the path between the calling party's and called party's locations from the global topology that it has in memory from LBM replication. If it is successful, Cluster 2 will replicate the reservation and extend the call to the terminating device and return a 180 ringing with the location information of the called party back to Cluster 1. When Cluster 1 receives the 180 ringing, it gets the terminating device's location name and goes through the same bandwidth lookup process using the same unique call-ID that it calculates from the information passed in the call-info header. If it is successful, it too continues with the call flow. Because both clusters use the same information in the call-info header, they will deduct bandwidth for the same call using the same call-ID, thus avoiding any double bandwidth deductions.

Location and Link Management Cluster

In order to avoid configuration overhead, a Location and Link Management Cluster can be configured to manage all locations and links in the global topology. All other clusters uniquely configure the locations that they require for location-to-device association and do not configure links or any bandwidth values other than unlimited. It should be noted that the Location and Link Management Cluster is a design concept and is simply any cluster that is configured with the entire global topology of locations and links, while all other clusters in the LBM replication network are configured only with locations set to unlimited bandwidth values and without configured links. When intercluster Enhanced Location CAC is enabled and the LBM replication network is configured, all clusters replicate their view of the network. The designated Location and Link Management Cluster has the entire global topology with

locations, links, and bandwidth values; and once those values are replicated, all clusters use those values because they are the most restrictive. This design alleviates configuration overhead in deployments where a large number of common locations are required across multiple clusters.

Recommendations

- Locations and link management cluster:
  - One cluster should be chosen as the management cluster (the cluster chosen to manage locations and links).
  - The management cluster should be configured with the following:

    All locations within the enterprise will be configured in this cluster.

    All bandwidth values and weights for all locations and links will be managed in this cluster.
- All other clusters in the enterprise:
  - All other clusters in the enterprise should configure *only* the locations required for association to devices but should *not* configure the links between locations. This link information will come from the management cluster when intercluster Enhanced Location CAC is enabled.
  - When intercluster Enhanced Location CAC is enabled, all of the locations and links will be replicated from the management cluster and LBM will use the lowest, most restrictive bandwidth and weight value.
- LBM will always use the lowest most restrictive bandwidth and lowest weight value after replication.

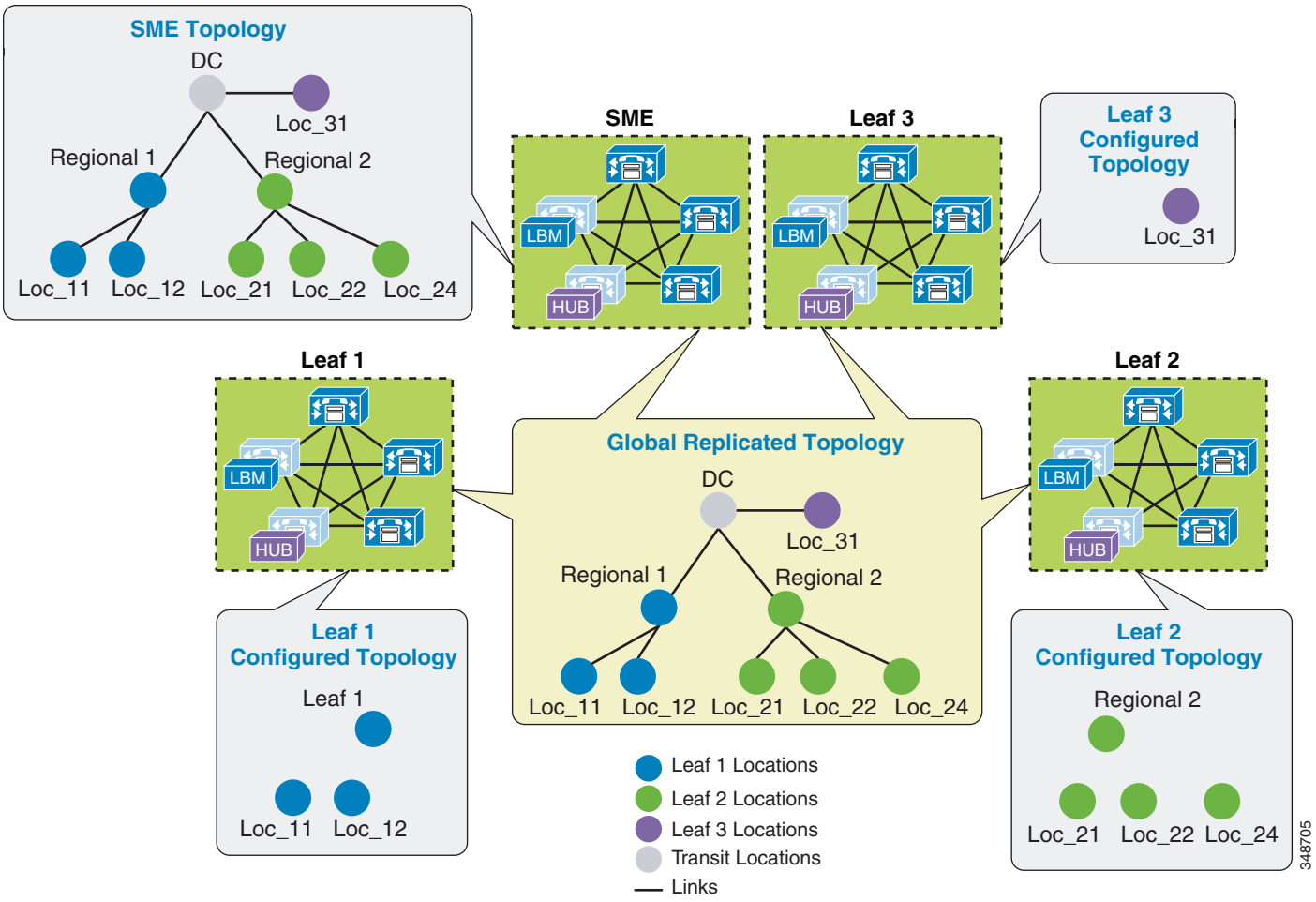
Benefits

- Manage enterprise CAC topology from a single cluster.
- Alleviates location and link configuration overhead when clusters share a large number of common locations.
- Alleviates configuration mistakes in locations and links across clusters.
- Other clusters in the enterprise require the configuration only of locations needed for location-to-device and endpoint association.
- Provides a single cluster for monitoring of the global locations topology.

Figure 13-35 illustrates Cisco Unified Communications Manager Session Management Edition (SME) as a Location and Link Management Cluster for three leaf clusters.

Note As mentioned, any cluster can act as the Location and Link management cluster. In this example SME is the Location and Link management cluster.

Figure 13-35 Example of SME as a Location and Link Management Cluster



In Figure 13-35 there are three leaf clusters, each with devices in only a regional and remote locations. SME has the entire global topology configured with locations and links, and intercluster LBM replication is enabled between all four clusters. None of the clusters in this example share locations, although all of the locations are common locations because SME has configured the entire location and link topology. Note that Leaf 1, Leaf 2, and Leaf 3 configure only locations that they require to associate to devices and endpoints, while SME has the entire global topology configured. After intercluster replication, all clusters will have the global topology.

Intercluster Enhanced Location CAC Design and Deployment Recommendations and Considerations

- A cluster requires the location to be configured locally for location-to-device association.
- Each cluster should be configured with the immediately neighboring locations so that each cluster's topology can inter-connect. This does not apply to Location and Link Management Cluster deployments.
- Links need to be configured to establish points of interconnect between remote topologies. This does not apply to Location and Link Management Cluster deployments.

- Discrepancies of bandwidth limits and weights on common locations and links are resolved by using the lowest bandwidth and weight values.
- Naming locations consistently across clusters is critical. Follow the practice, "Same location, same name; different location, different name."
- The Hub_None location should be renamed to be unique in each cluster or else it will be a common (shared) location by other clusters.
- Cluster-ID should be unique on each cluster for serviceability reports to be usable.
- All LBM hubs are fully meshed between clusters.
- An LBM hub is responsible for communicating to hubs in remote clusters.
- An LBM spoke does not directly communicate with other remote clusters. LBM spokes receive and send messages to remote clusters through the Local LBM Hub.
- LBM Hub Groups
  - Used to assign LBMs to the Hub role
  - Used to define three remote hub members that replicate hub contact information for all of the hubs in the LBM hub replication network
  - An LBM is a hub when it is assigned to an LBM hub group.
  - An LBM is a spoke when it is not assigned to an LBM hub group.
- If a cluster has no LBM hub, or if the LBM hub is not running, the cluster will be isolated and will not participate in the intercluster LBM replication network.

Performance Guidelines

- Maximum of 2,000 locally configured locations. This limit of 2,000 locations also applies to the Location and Link Management Cluster.
- Maximum of 8,000 total replicated locations with intercluster CAC

Enhanced Location CAC for TelePresence Immersive Video

Since TelePresence endpoints now provide a diverse range of collaborative experiences from the desktop to the conference room, Enhanced Location CAC includes support to provide CAC for TelePresence immersive video calls. This section discusses the features in Enhanced Location CAC that support TelePresence immersive video CAC.

Video Call Traffic Class

Video Call Traffic Class is an attribute that is assigned to all endpoints, and that can also be enabled on SIP trunks, to determine the video classification type of the endpoint or trunk. This enables Unified CM to classify various call flows as either immersive, desktop video, or both, and to deduct accordingly from the appropriate location and/or link bandwidth allocations of video bandwidth, immersive bandwidth, or both. For TelePresence endpoints there is a non-configurable Video Call Traffic Class of **immersive** assigned to the endpoint. A SIP trunk can be classified as desktop, immersive, or mixed video in order to deduct bandwidth reservations of a SIP trunk call. All other endpoints and trunks have a non-configurable Video Call Traffic Class of **desktop video**. More detail on endpoint and trunk classification is provided in the subsections below.

TelePresence immersive endpoints mark their media with a DSCP value of CS4 by default, and desktop video endpoints mark their media with AF41 by default, as per recommended QoS settings. For Cisco endpoints this is accomplished through the configurable Unified CM QoS service parameters **DSCP for Video calls** and **DSCP for TelePresence calls**. When a Cisco TelePresence endpoint registers with Unified CM, it downloads a configuration file and applies the QoS setting of **DSCP for TelePresence calls**. When a Unified Communications video-capable endpoint registers with Unified CM, it downloads a configuration file and applies the QoS setting of **DSCP for Video calls**. All third-party video endpoints require manual configuration of the endpoints themselves and are statically configured, meaning they do not change QoS marking depending on the call type; therefore, it is important to match the Enhanced Location CAC bandwidth allocation to the correct DSCP. Unified CM achieves this by deducting desktop video calls from the Video Bandwidth location and link allocation for devices that have a Video Call Traffic Class of **desktop**. End-to-end TelePresence immersive video calls are deducted from the Immersive Video Bandwidth location and link allocation for devices or trunks with the Video Call Traffic Class of **immersive**. This ensures that end-to-end desktop video and immersive video calls are marked correctly and counted correctly for call admission control. For calls between desktop devices and TelePresence immersive devices, bandwidth is deducted from both the Video Bandwidth and the Immersive Video Bandwidth location and link allocations.

Endpoint Classification

Cisco TelePresence endpoints have a fixed non-configurable Video Call Traffic Class of **immersive** and are identified by Unified CM as immersive. Telepresence endpoints are defined in Unified CM by the device type. When a device is added in Unified CM, any device with TelePresence in the name of the device type is classified as **immersive**, as are the generic single-screen and multi-screen room systems. Another way to check the capabilities of the endpoints in the Unified CM is to go to the **Cisco Unified Reporting Tool > System Reports > Unified CM Phone Feature List**. In the feature drop down list, select **Immersive Video Support for TelePresence Devices**; in the product drop down list, select **All**. This will display all of the device types that are classified as **immersive**. All other endpoints have a fixed Video Call Traffic Class of **desktop** due to their lack of the non-configurable immersive attribute.

Bandwidth reservations are determined by the classification of endpoints in a video call, and they deduct bandwidth from the locations and links bandwidth pools as listed in Table 13-12.

Table 13-12 Bandwidth Pool Usage per Endpoint Type

Endpoint A	Endpoint B	Locations and Links Pool Used
Immersive video	Immersive video	Immersive bandwidth
Immersive video	Desktop video	Immersive and video bandwidth
Desktop video	Desktop video	Video bandwidth
Audio-only call	Any	Audio bandwidth

SIP Trunk Classification

A SIP trunk can also be classified as desktop, immersive, or mixed video in order to deduct bandwidth reservations of a SIP trunk call, and the classification is determined by the calling device type and Video Call Traffic Class of the SIP trunk. The SIP trunk can be configured through the SIP Profile trunk-specific information as:

- Immersive — High-definition immersive video
- Desktop — Standard desktop video
- Mixed — A mix of immersive and desktop video

A SIP trunk can be classified with any of these three classifications and is used primarily to classify Video or TelePresence Multipoint Control Units (MCUs), a video device at a fixed location, a Unified CM cluster supporting traditional Location CAC, or a Cisco TelePresence System Video Communications Server (VCS).

Bandwidth reservations are determined by the classification of an endpoint and a SIP trunk in a video call, and they deduct bandwidth from the locations and links bandwidth pools as listed in Table 13-13.

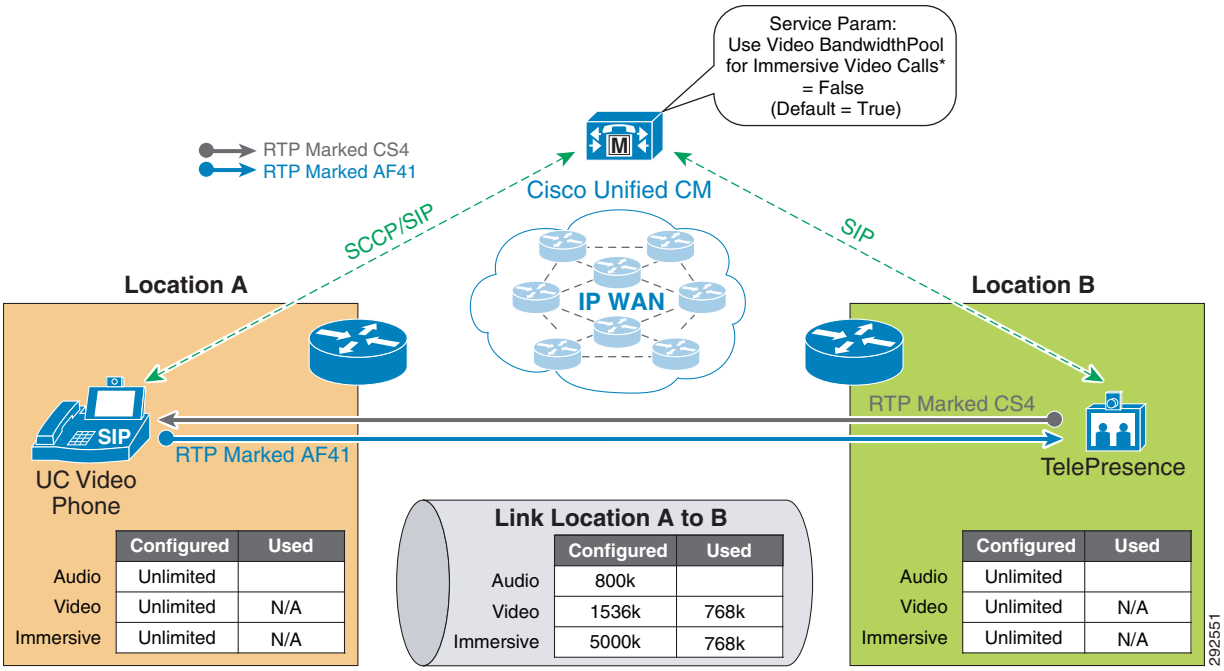
Table 13-13 Bandwidth Pool Usage per SIP Trunk and Endpoint Type

Endpoint	SIP Trunk	Locations and Links Pool Used
TelePresence endpoint	Immersive	Immersive bandwidth
TelePresence endpoint	Desktop	Immersive and video bandwidth
TelePresence endpoint	Mixed	Immersive and video bandwidth
Desktop endpoint	Immersive	Immersive and video bandwidth
Desktop endpoint	Desktop	Video bandwidth
Desktop endpoint	Mixed	Immersive and video bandwidth
Non-video endpoint	Any	Audio bandwidth

By default, all video calls from either immersive or desktop endpoints are deducted from the locations and links video bandwidth pool. To change this behavior, set Unified CM's CallManager service parameter **Use Video BandwidthPool for Immersive Video Calls** to **False**, and this will enable the immersive video bandwidth deductions. After this is enabled, immersive and desktop video calls will be deducted out of their respective pools.

As described earlier, a video call between a Unified Communications video endpoint (desktop Video Call Traffic Class) and a TelePresence endpoint (immersive Video Call Traffic Class) will mark their media asymmetrically and, when immersive video CAC is enabled, will deduct bandwidth from both video and immersive locations and links bandwidth pools. Figure 13-36 illustrates this.

Figure 13-36 Enhanced Location CAC Bandwidth Deductions and Media Marking for a Multi-Site Deployment



Examples of Various Call Flows and Location and Link Bandwidth Pool Deductions

The following call flows depict the expected behavior of locations and links bandwidth deductions when the Unified CM service parameter **Use Video BandwidthPool for Immersive Video Calls** is set to **False**.

Figure 13-37 illustrates an end-to-end TelePresence immersive video call between TP-A in Location L1 and TP-B in Location L2. End-to-end immersive video endpoint calls deduct bandwidth from the immersive bandwidth pool of the locations and the links along the effective path.

Figure 13-37 Call Flow for End-to-End TelePresence Immersive Video

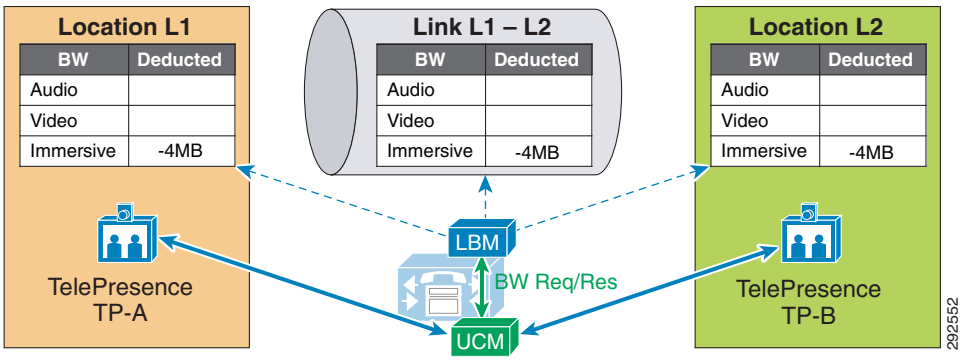


Figure 13-38 illustrates an end-to-end desktop video call between DP-A in Location L1 and DP-B in Location L2. End-to-end desktop video endpoint calls deduct bandwidth from the video bandwidth pool of the locations and the links along the effective path.

Figure 13-38 Call Flow for End-to-End Desktop Video

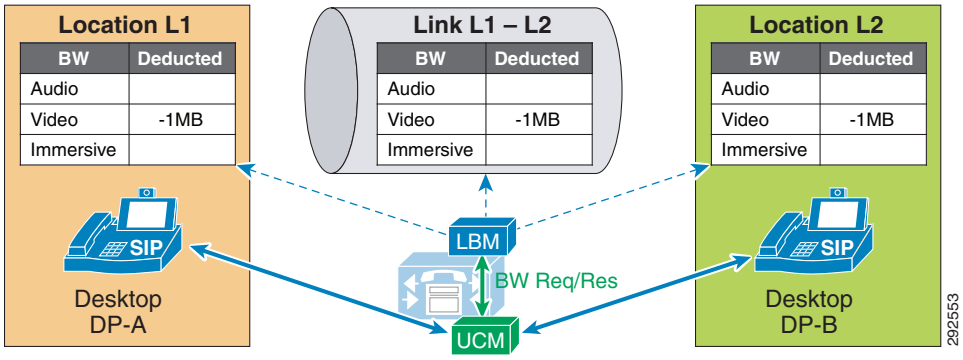
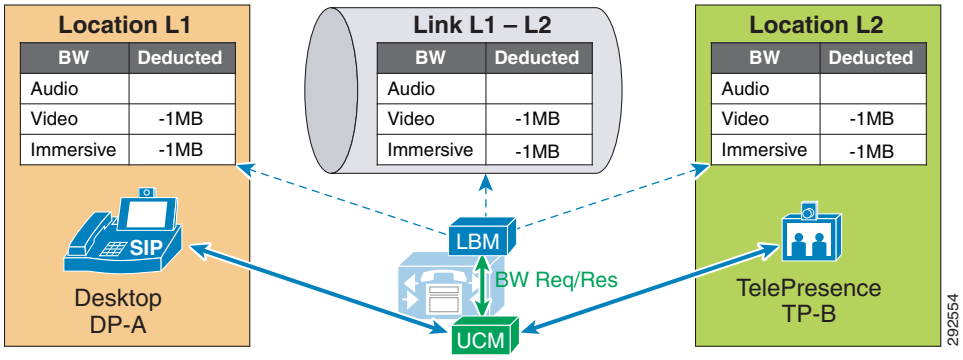


Figure 13-39 illustrates a video call between desktop video endpoint DP-A in Location L1 and TelePresence video endpoint TP-B in Location L2. Interoperating calls between desktop video endpoints and TelePresence video endpoints deduct bandwidth from both video and immersive locations and the links bandwidth pools along the effective path.

Figure 13-39 Call Flow for Desktop-to-TelePresence Video



In Figure 13-40, a desktop video endpoint and two TelePresence endpoints call a SIP trunk configured with a Video Traffic Class of **immersive** that points to a TelePresence MCU. Bandwidth is deducted along the effective path from the immersive locations and the links bandwidth pools for the calls that are end-to-end immersive and from both video and immersive locations and the links bandwidth pools for the call that is desktop-to-immersive.

Figure 13-40 Call Flow for a Video Conference with an MCU

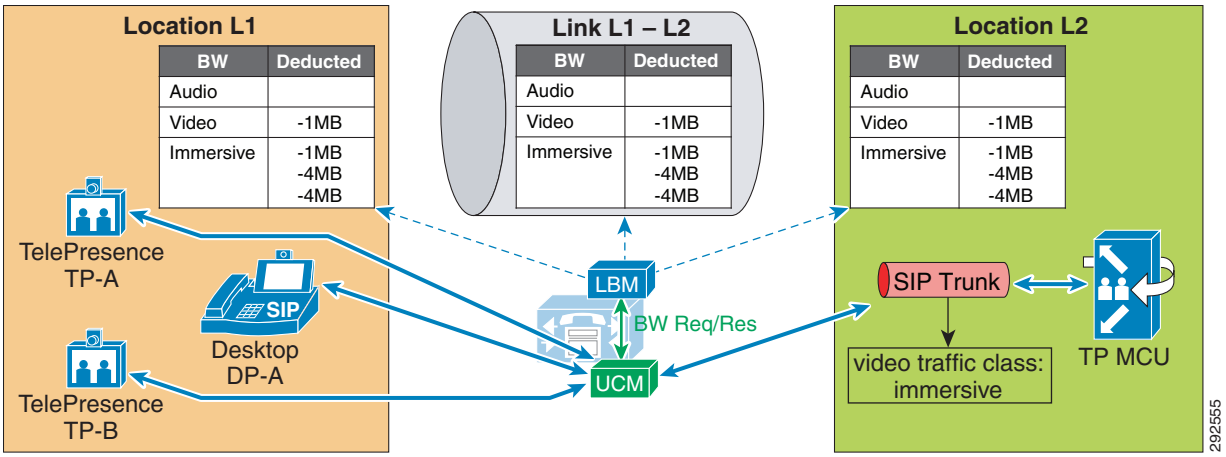


Figure 13-41 illustrates an end-to-end immersive video call across clusters, which deducts bandwidth from the immersive bandwidth pool of the locations and links along the effective path.

Figure 13-41 Call Flow for End-to-End TelePresence Immersive Video Across Clusters

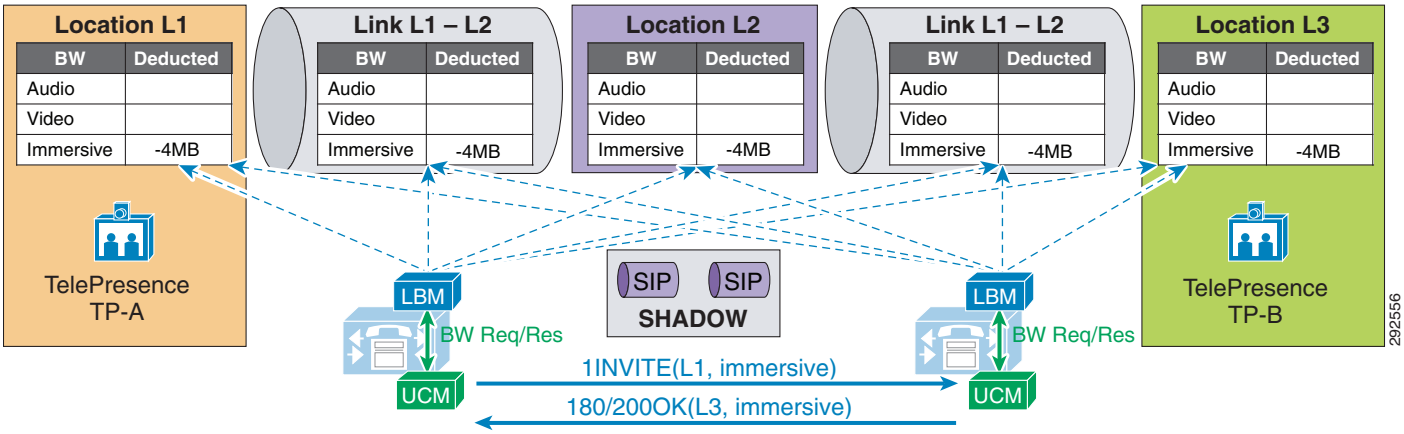


Figure 13-42 illustrates an end-to-end desktop video call across clusters, which deducts bandwidth from the video bandwidth pool of the locations and links along the effective path.

Figure 13-42 Call Flow for End-to-End Desktop Video Call Across Clusters

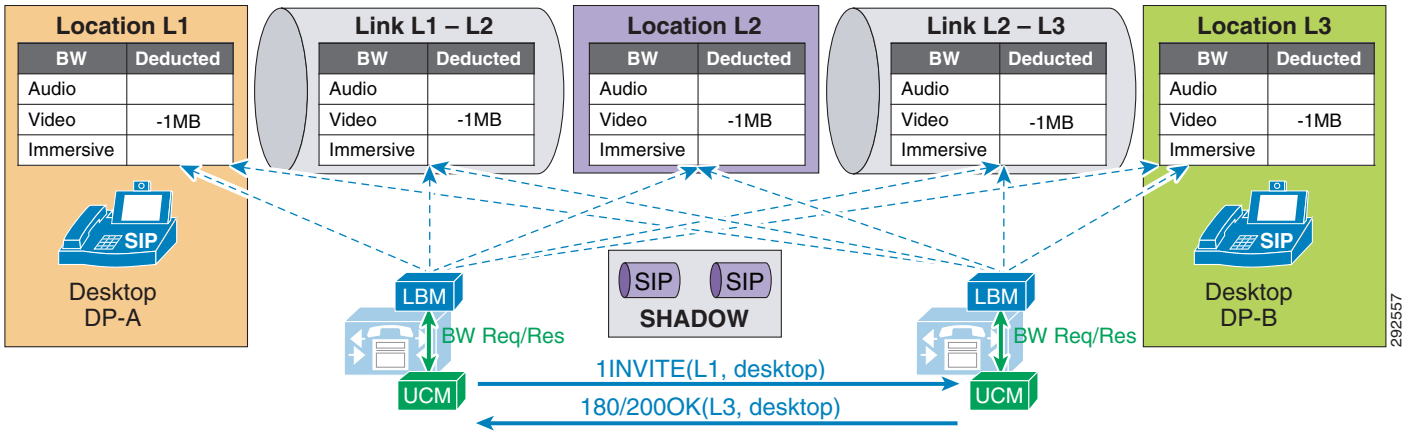
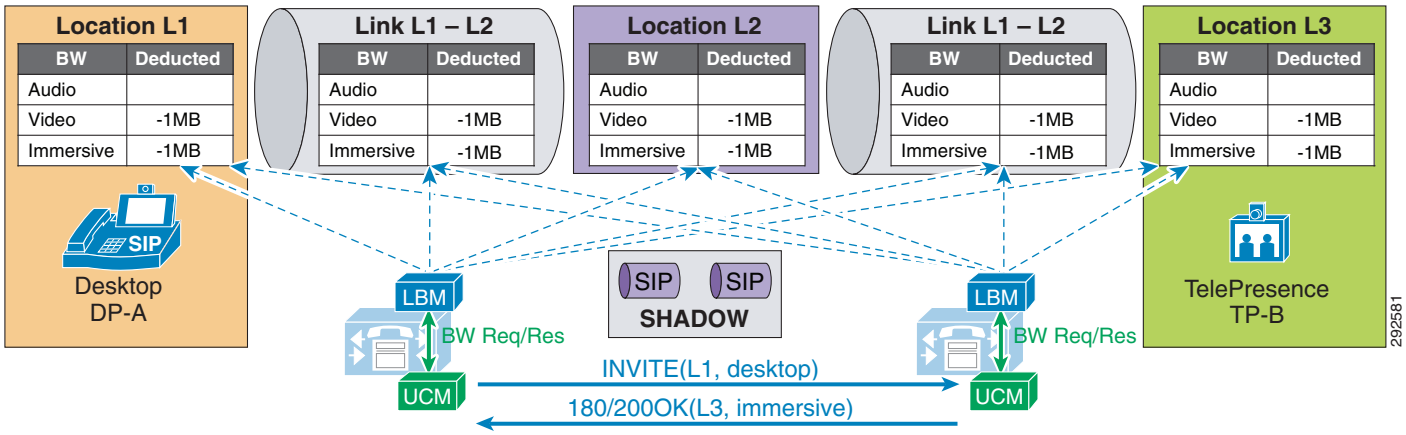


Figure 13-43 illustrates a desktop video endpoint calling a TelePresence endpoint across clusters. the call deducts bandwidth from both video and immersive bandwidth pools of the locations and links along the effective path.

Figure 13-43 Call Flow for Desktop-to-TelePresence Video Across Clusters



Video Bandwidth Utilization and Admission Control

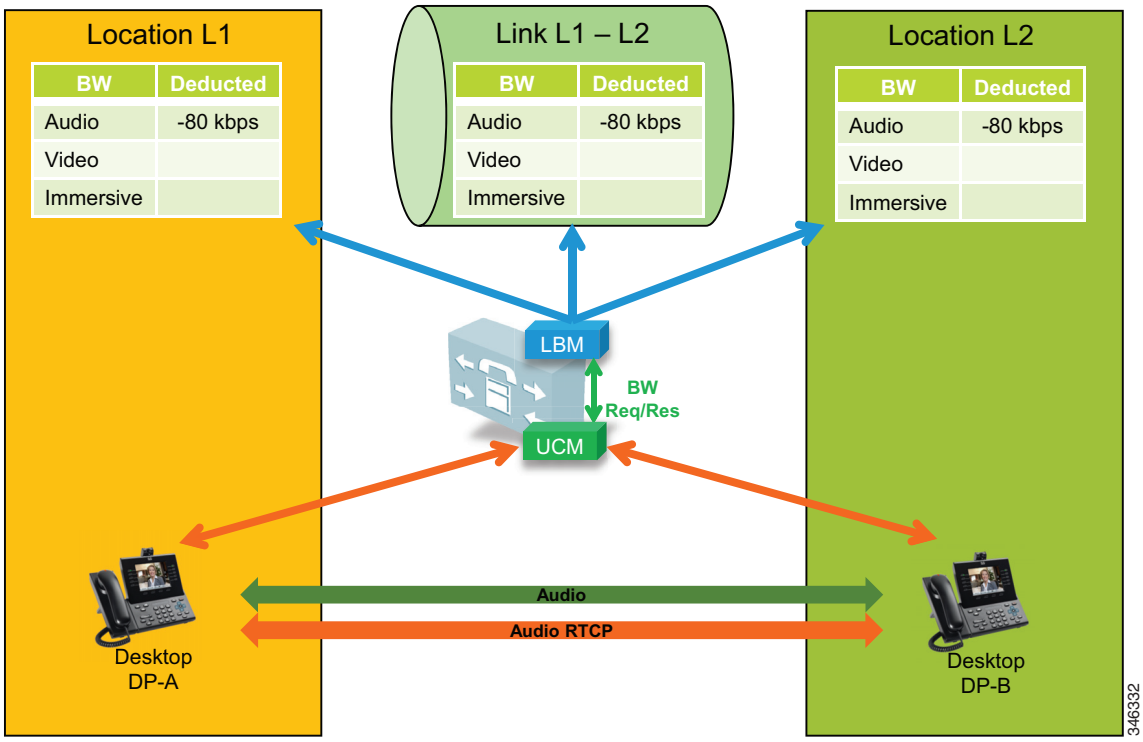
When Unified CM negotiates an audio or video call, a number of separate streams are established between the endpoints involved in the call. For video calls with content sharing, this can result in as many as 8 (or possibly more) unidirectional streams. For an audio-only call typically the bare minimum is 2 streams, one in each direction. This section discusses bandwidth utilization on the network and how Unified CM accounts for this in admission control bandwidth accounting.

For the purpose of the discussion in this section, please note the following:

- The figures in this section use a bidirectional arrow (<-->) to represent two unidirectional streams.
- The following points summarize how Unified CM Enhanced Location CAC deducts bandwidth from the configured audio, video, and immersive allocations. For more information, see the section on Locations and Links, page 13-42:
  - Audio (audio-only calls): RTP bit rate + IP and UDP header overhead
  - Video (video calls): RTP bit rate only
  - Immersive (video calls by Cisco TelePresence endpoints): RTP bit rate only
- Bandwidth deductions in Enhanced Location CAC:
  - Bandwidth deductions are made for bidirectional RTP streams and are assumed to be symmetrically routed (both streams routed over the same path). For example, a G.711 audio call of 80 kbps is 80 kbps in each direction over a full-duplex network; that is 80 kbps on the transmit pair of wires and 80 kbps on the receive pair of wires, equating to 80 kbps full-duplex. (See Figure 13-44.) Note that traffic is not always routed symmetrically in the WAN. Check with your network administrator when necessary to ensure that admission control is correctly accounting for the media as it is routed in the network over the WAN.
  - Real-Time Transport Control Protocol (RTCP) bandwidth overhead is not part of Unified CM bandwidth allocation and should be part of network provisioning. RTCP is quite common in most call flows and is commonly used for statistical information about the streams. It is also used to synchronize audio in video calls to ensure proper lip-sync. In some cases it can be enabled or disabled on the endpoint. RFC 3550 recommends that the fraction of the session bandwidth added for RTCP should be fixed at 5%. What this means is that it is common practice for the RTCP session to be up to 5% of the associated RTP session. So when calculating bandwidth consumption on the network, you should add the RTCP overhead for each RTP session. For example, if you have a G.711 audio call of 80 kbps with RTCP enabled, you will be using up to 84 kbps per session (4 kbps RTCP overhead) for both RTP and RTCP. This calculation is not part of Enhanced Location CAC deductions but should be part of network provisioning.

Note There are, however, methods to re-mark this traffic to another Differentiated Services Code Point (DSCP). For example, RTCP uses odd-numbered UDP ports while RTP uses even-numbered UDP ports. Therefore, classification based on UDP port ranges is possible. Network-Based Application Recognition (NBAR) is another option as it allows for classification and re-marking based on the RTP header **Payload Type** field. For more information on NBAR, see https://www.cisco.com. However, if the endpoint marking is trusted in the network, then RTCP overhead should be provisioned in the network within the same QoS class as audio RTP (default marking is EF). It should also be noted that RTCP is marked by the endpoint with the same marking as RTP; by default this is EF (verify that RTCP is also marked as EF).

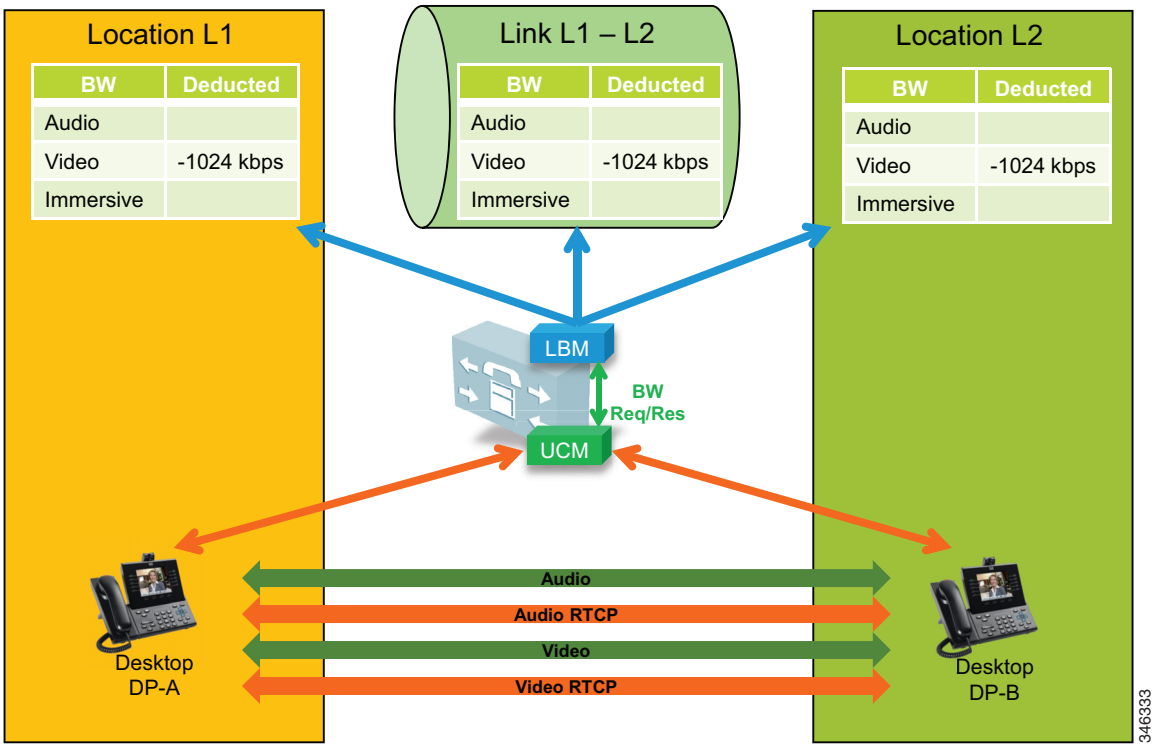
Figure 13-44 A Basic Audio-Only Call with RTCP Enabled



In Figure 13-44 two desktop video phones have established an audio-only call. In this call flow four streams are negotiated: two audio streams illustrated by a single bidirectional arrow and two RTCP streams also illustrated by a bidirectional arrow. For this call, the Location Bandwidth Manager (LBM) deducts 80 kbps (bit rate + IP/UDP overhead) between location L1 and location L2 for a call established between desktop phones DP-A and DP-B. The actual bandwidth consumed at Layer 3 in the network with RTCP enabled would be between 80 kbps and 84 kbps, as discussed previously in this section.

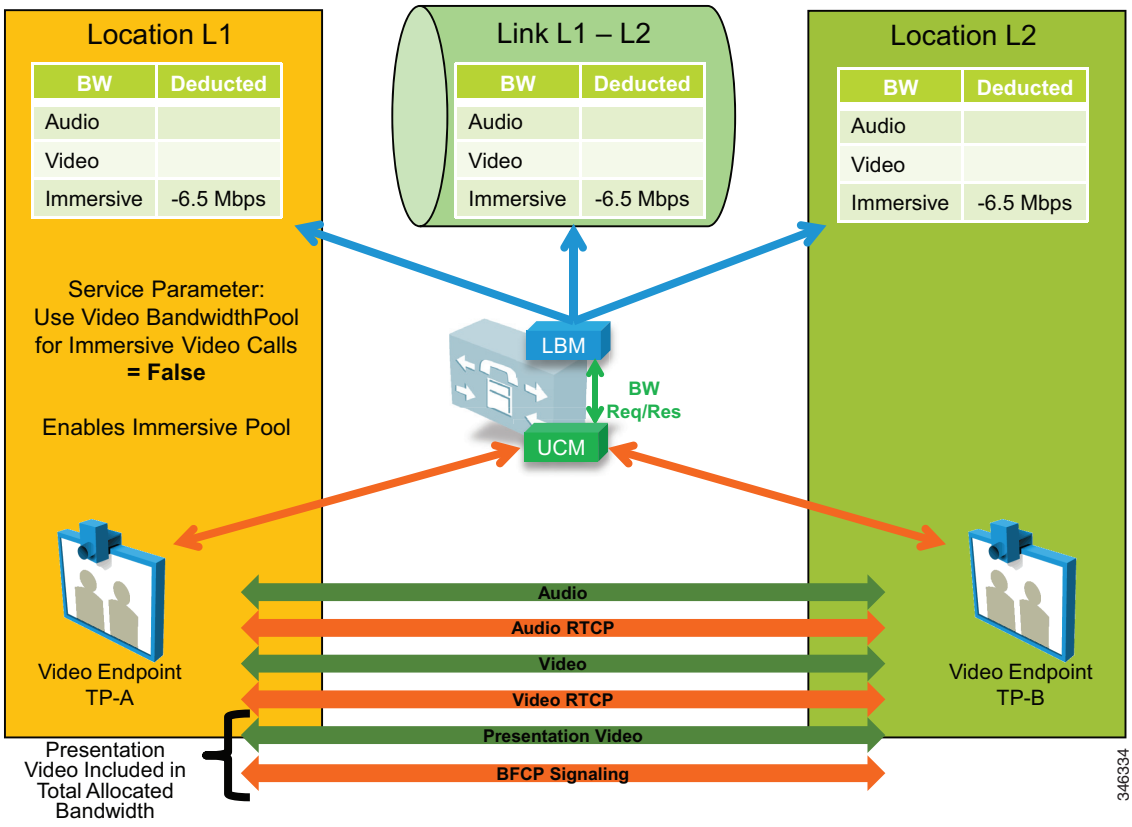
In Figure 13-45 two desktop video phones have established a video call. In this call flow eight streams are negotiated: two audio streams, two audio-associated RTCP streams, two video streams, and two video-associated RTCP streams. Again for this illustration one bidirectional arrow is used to depict two unidirectional streams. This particular call is 1024 kbps, with 64 kbps of G.711 audio and 960 kbps of video (bit rate only for audio and video allocations of video calls). So in this case the LBM deducts 1024 kbps between locations L1 and L2 for a video call established between desktop phones DP-A and DP-B. RTCP is overhead that should be accounted for in provisioning, depending on how it is marked or re-marked by the network.

Figure 13-45 A Basic Video Call with RTCP Enabled



The example in Figure 13-46 is of a video call with presentation sharing. This is a more complex call with regard to the number of associated streams and bandwidth allocation when compared to bandwidth used on the network, and therefore it must be provisioned in the network. Figure 13-46 illustrates a video call with RTCP enabled and Binary Floor Control Protocol (BFCP) enabled for presentation sharing. All SIP-enabled telepresence multipurpose or personal endpoints such as a the Cisco TelePresence System EX, MX, SX, C, and Profile Series function in the same manner.

Figure 13-46 Video Call with RTCP and BFCP Enabled and Presentation Sharing

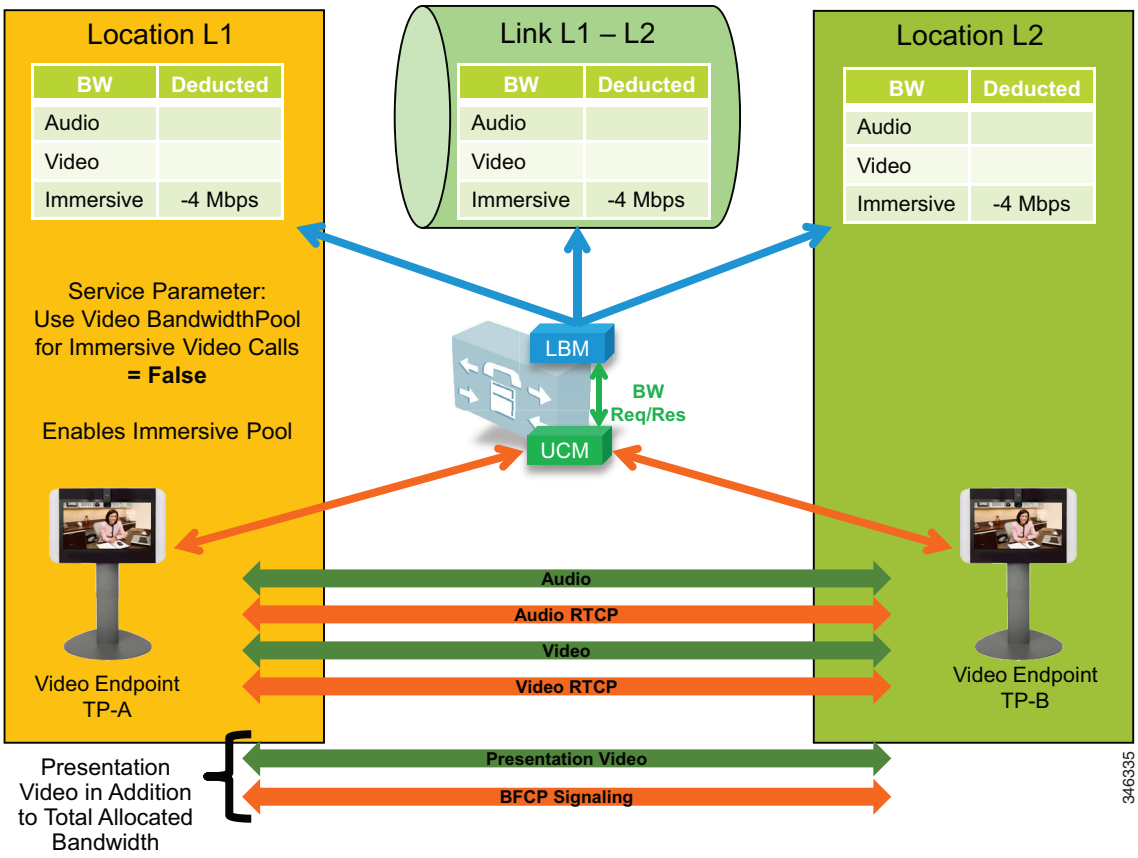


When a video call is established between two video endpoints, audio and video streams are established and bandwidth is deducted for the negotiated rate. Unified CM uses regions to determine the maximum bit rate for the call. For example, with a Cisco TelePresence System EX90 at the highest detail of 1080p at 30 frames per second (fps), the negotiated rate between regions would have to be set at 6.5 Mbps. EX90s used in this scenario would average around 6.1 Mbps for this session. When the endpoints start presentation sharing during the session, BFCP is negotiated between the endpoints and a new video stream is enabled at either 5 fps or 30 fps, depending on endpoint configuration. When this occurs, the endpoints will throttle down their main video stream to include the presentation video so that the entire session does not use more than the allocated bandwidth of 6.5 Mbps. Thus, the average bandwidth consumption remains the same with or without presentation sharing.

Note The presentation video stream is typically unidirectional in the direction of the person or persons viewing the presentation.

Telepresence immersive and office endpoints such as the Cisco TelePresence System 500, 1000, 3000, and TX9000 Series that negotiate a call between one another function a little differently in the sense that the video for presentation sharing is an additional bandwidth above and beyond what is allocated for the main video session, and thus it is not deducted from Enhanced Location CAC. Figure 13-47 illustrates this.

Figure 13-47 Video Call with RTCP and BFCP Enabled and Presentation



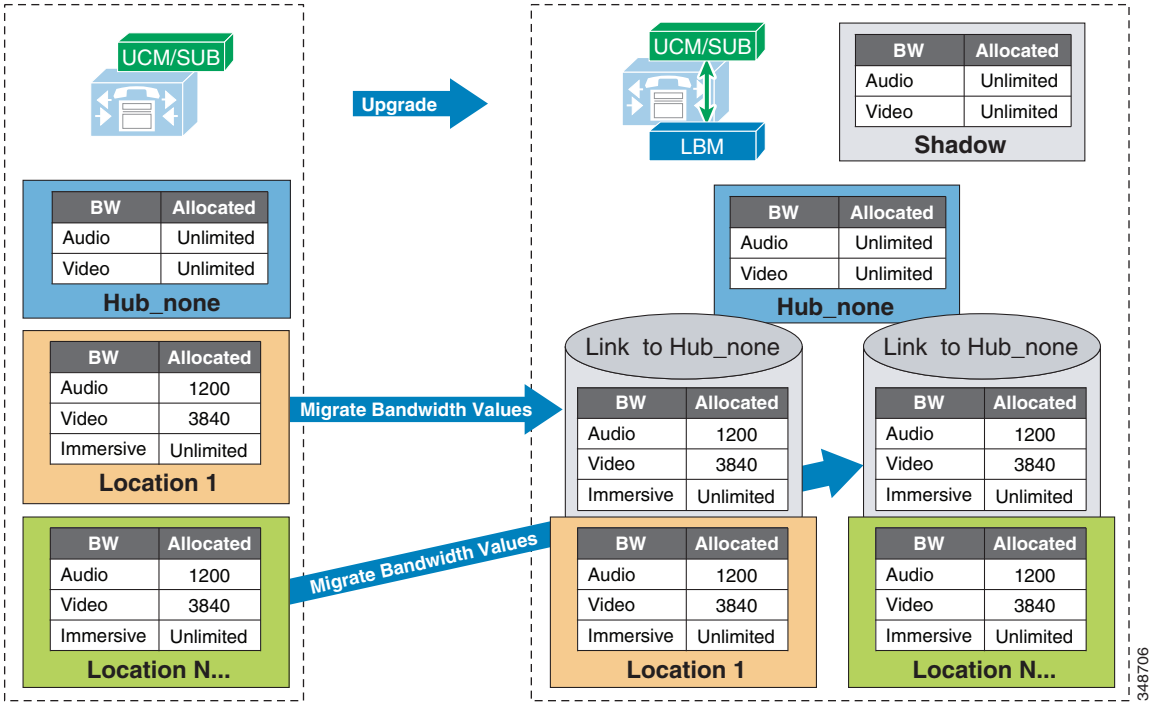
In Figure 13-47 the telepresence immersive video endpoints establish a video call and enable presentation sharing. The LBM deducts 4 Mbps for the main audio and video session from the immersive pool for the call, and video is established between the endpoints. When presentation sharing is activated, the two endpoints exchange BFCP and negotiate a presentation video stream at 5 fps or 30 fps in one direction, depending on the endpoint configuration. At 5 fps the average bandwidth used is approximately 500 kbps of additional bandwidth overhead. This bandwidth is above and beyond the 4 Mbps that was allocated for the video call and should be provisioned in the network. At 30 fps the average bit rate of the presentation video is approximately 1.5 Mbps.

Note The Cisco TelePresence System endpoints use Telepresence Interoperability Protocol (TIP) to multiplex multiple screens and audio into two audio and video RTP streams in each direction. Therefore the actual streams on the wire may be different than what is expressed in the illustration, but the concept of additional bandwidth overhead for the presentation video is the same.

Upgrade and Migration from Location CAC to Enhanced Location CAC

Upgrading to Cisco Unified CM from a previous release that supports only traditional Location CAC, will result in the migration of Location CAC to Enhanced Location CAC. The migration consists of taking all previously defined locations bandwidth limits of audio and video bandwidth and migrating them to a link between the user-defined location and Hub_None. This effectively recreates the hub-and-spoke model that previous versions of Unified CM Location CAC supported. Figure 13-48 illustrates the migration of bandwidth information.

Figure 13-48 Migration from Location CAC to Enhanced Location CAC After Unified CM Upgrade



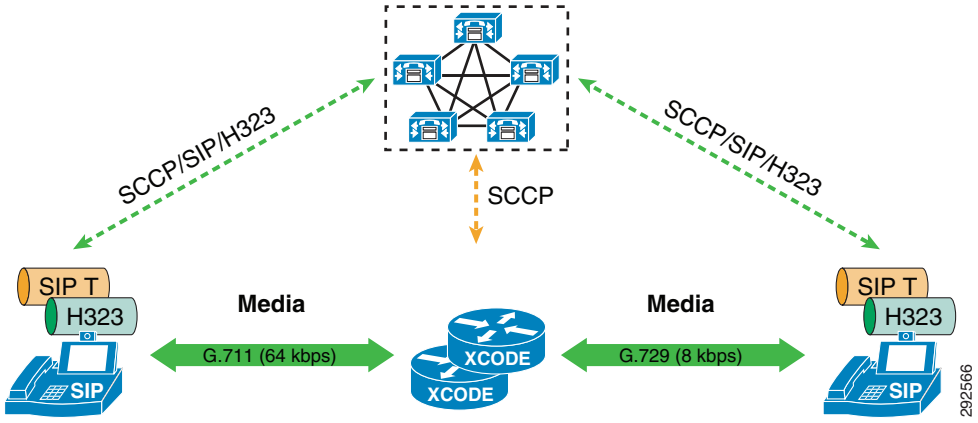
Settings after an upgrade to a Cisco Unified CM release that supports Enhanced Location CAC:

- The LBM is activated on each Unified CM subscriber running the Cisco CallManager service.
- The Cisco CallManager service communicates directly with the local LBM.
- No LBM group or LBM hub group is created.
- All LBM services are fully meshed.
- Intercluster Enhanced Location CAC is not enabled.
- All intra-location bandwidth values are set to unlimited.
- Bandwidth values assigned to locations are migrated to a link connecting the user-defined location and Hub_None.
- Immersive bandwidth is set to unlimited.
- A shadow location is created.

- Phantom and shadow locations have no links.
- Enhanced Location CAC bandwidth adjustment for MTPs and transcoders:

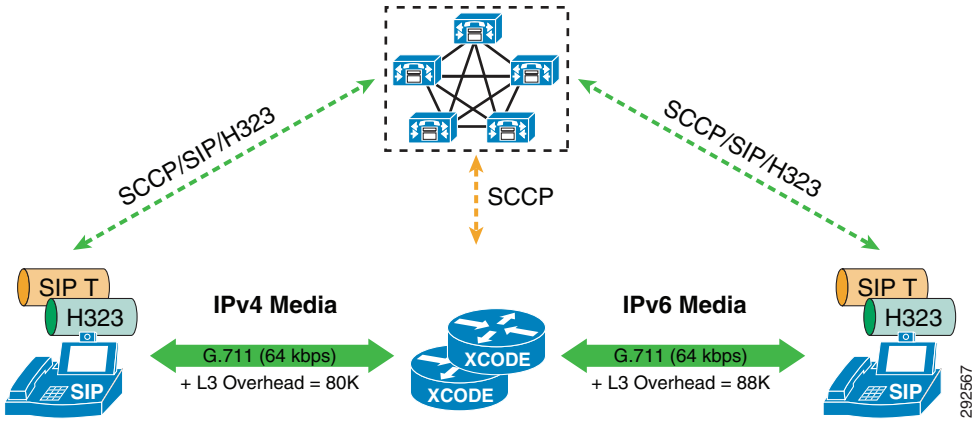
  For transcoding insertion, the bit rate is different on each leg of the connection. Figure 13-49 illustrates this.

Figure 13-49 Example of Different Bit Rate for Transcoding



For dual stack MTP insertion, the bit rate is different on each connection but the bandwidth is different due to IP header overhead. Figure 13-50 illustrates the difference in bandwidth used for IPv4 and IPv6 networks with dual stack MTP insertion.

Figure 13-50 Bandwidth Differences for Dual Stack MTP Insertion



Enhanced Location CAC does not account for these differences in bandwidth between MTPs and transcoders. The service parameter **Locations Media Resource Audio Bit Rate Policy** determines whether the largest or smallest bandwidths should be used along the locations and links path. Lowest Bit Rate (default) or Highest Bit Rate can be used to manage these differences in bandwidth consumption.

Extension Mobility Cross Cluster with Enhanced Location CAC

Enhanced Location CAC supports designs using Extension Mobility Cross Cluster (EMCC). Unified CM provides the ability to perform Extension Mobility logins between clusters within an enterprise with a feature called Extension Mobility Cross Cluster (EMCC). For further information, see the section on Extension Mobility Cross Cluster (EMCC), page 18-9.

With Enhanced Location CAC in EMCC designs, the visiting cluster passes the location of the visiting phone to the home cluster. This allows the home cluster to associate the correct location to the visiting phone during registration. The following requirements must be met for Enhanced Location CAC to function in EMCC designs:

- Cisco Unified CM 10.0 or a later release required on both home and visiting clusters
- The visiting and home clusters must be in the same intercluster LBM replication network

Both Enhanced Location CAC and EMCC can be designed and deployed according to the guidelines in the product documentation and this SRND. There are no other requirements or any specific configuration aspects to employ.

Design Considerations for Call Admission Control

This section describes how to apply the call admission control mechanisms to various IP WAN topologies. With Unified CM Enhanced Location CAC network modeling support, Unified CM is no longer limited to supporting simple hub-and-spoke or MPLS topologies but, together with intercluster enhanced locations, can now support most any network topology in any Unified CM deployment model. Enhanced Location CAC is still a statically defined mechanism that does not query the network, and therefore the administrator still needs to provision Unified CM accordingly whenever network changes affect admission control. This is where a network-aware mechanism such as RSVP can fill that gap and provide support for dynamic changes in the network, such as when network failures occur and media streams take different paths in the network. This is often the case in designs with load-balanced dual or multi-homed WAN uplinks or unequally sized primary and backup WAN uplinks.

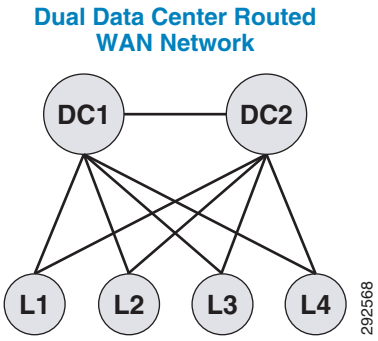
To learn how Enhanced Location CAC functions and how to design and deploy Enhanced location CAC, see the section on Unified CM Enhanced Location Call Admission Control, page 13-40.

In this section explores a few typical topologies and explains how Enhanced Location CAC can be designed to manage them.

Dual Data Center Design

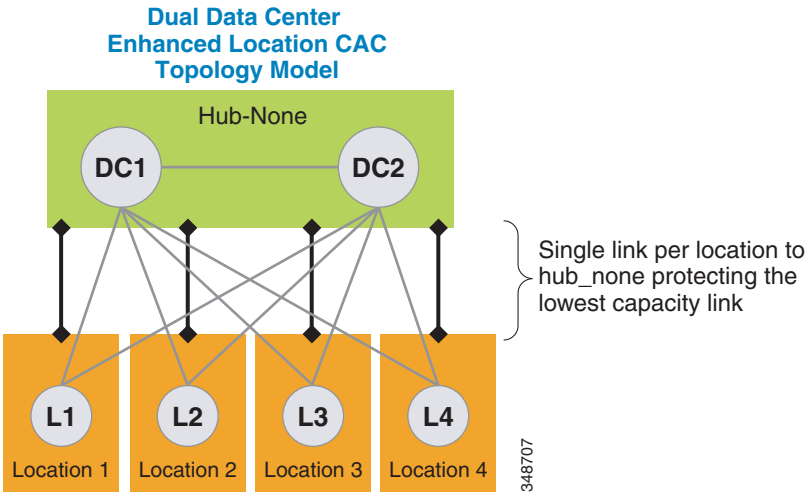
Figure 13-51 illustrates a simple dual data center WAN network design where each remote site has a single WAN uplink to each data center. The data centers are interconnected by a high-speed WAN connection that is over-provisioned for data traffic.

Figure 13-51 Dual Data Center WAN Network



Typically these WAN uplinks from the remote sites to the data centers are load-balanced or in a primary/backup configuration, and there are limited ways for a static CAC mechanism to handle these scenarios. Although you could configure this multi-path topology in Enhanced Location CAC, only one path would be calculated as the effective path and would remain statically so until the weight metric was changed. A better way to support this type of network topology is to configure the two data centers as one data center or hub location in Enhanced Location CAC and configure a single link to each remote site location. Figure 13-52 illustrates an Enhanced Location (E-L) CAC locations and links overlay.

Figure 13-52 Enhanced Location CAC Topology Model for Dual Data Centers



Design Recommendations

The following design recommendations for dual data centers with remote dual or more links to remote locations apply to both load-balanced and primary/backup WAN designs:

- A single location (Hub_None) represents both data centers.
- A single link between the remote locations and Hub_None protects the remote site uplinks from over-subscription during normal conditions or failure of the highest bandwidth capacity links.
- The capacity of link bandwidth allocation between the remote site and Hub_None should be equal to the lowest bandwidth capacity for the applicable Unified Communications media for a single link. For example, if each WAN uplink can support 2 Mbps of audio traffic marked EF, then the link audio bandwidth value should be no more than 2 Mbps to support a failure condition or equal-cost path routing.

MPLS Clouds

When designing for Multiprotocol Label Switching (MPLS) any-to-any connectivity type clouds in the Enhanced Location CAC network model, a single location can serve as the MPLS cloud. This location will not have any devices associated to it, but all of the sites that have uplinks to this cloud will have links configured to the location. In this way the MPLS cloud serves as a transit location for interconnecting multiple variable-sized bandwidth WAN uplinks to other remote locations. The illustrations in this section depict a number of different MPLS networks and their equivalent locations and links model.

In Figure 13-53, Hub_None represents the MPLS cloud serving as a transit location interconnecting the campus location where servers, endpoints, and devices are located, with remote locations where only endpoints and devices are located. Each link to Hub_None from the remote location may be sized according to the WAN uplink bandwidth allocated for audio, video, and immersive media.

Figure 13-53 Single MPLS Cloud

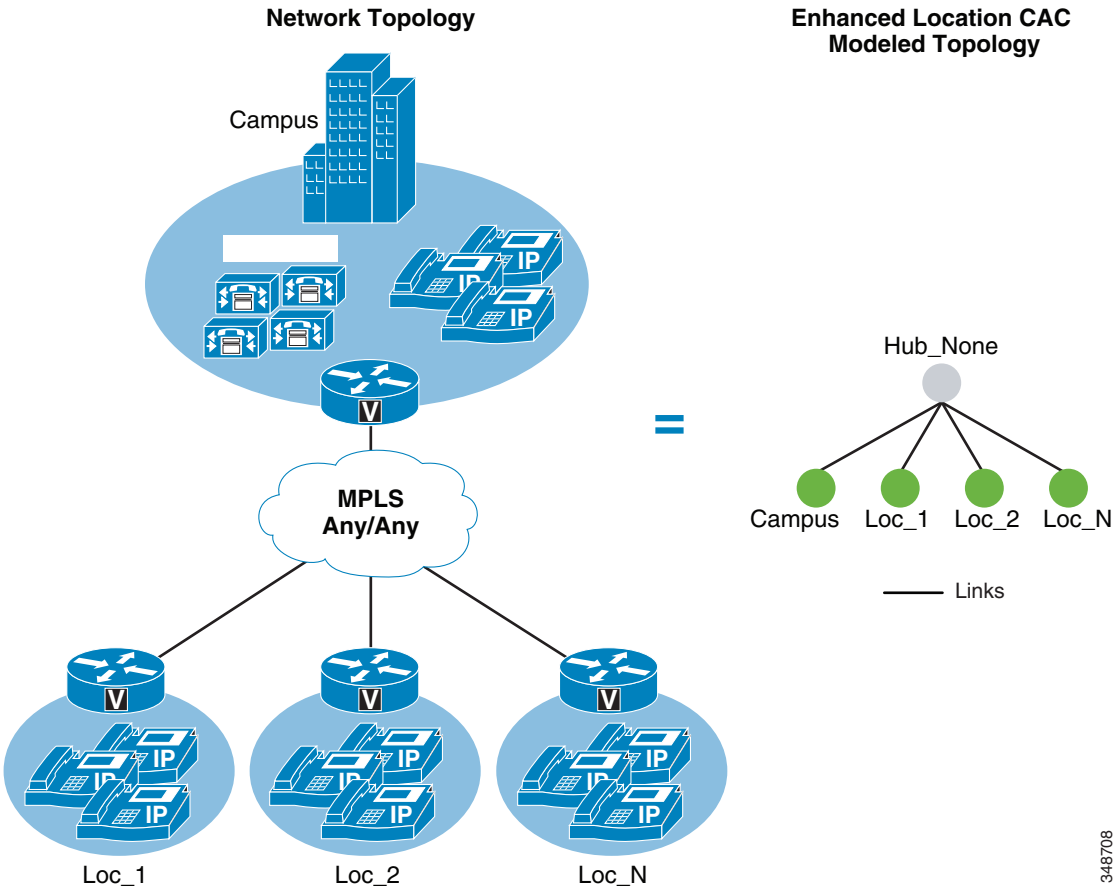


Figure 13-54 shows two MPLS clouds that serve as transit locations interconnecting the campus location where servers, endpoints, and devices are located, with remote locations where only endpoints and devices are located. The campus also connects to both clouds. Each link to the MPLS cloud from the remote location may be sized according to the WAN uplink bandwidth allocated for audio, vide, and immersive media. This design is typical in enterprises that span continents, with a separate MPLS cloud from different providers in each geographical location.

Figure 13-54 Separate MPLS Clouds

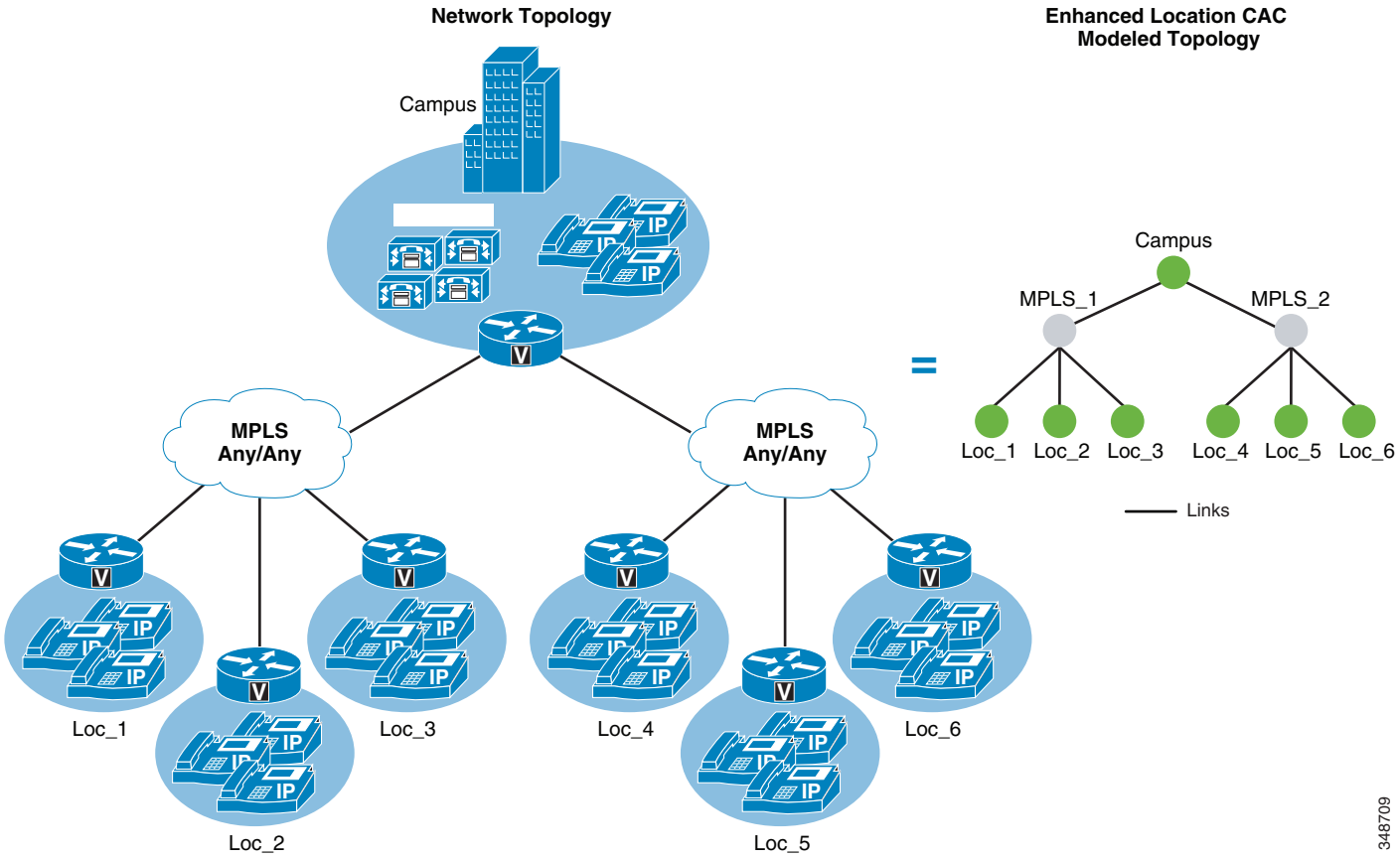
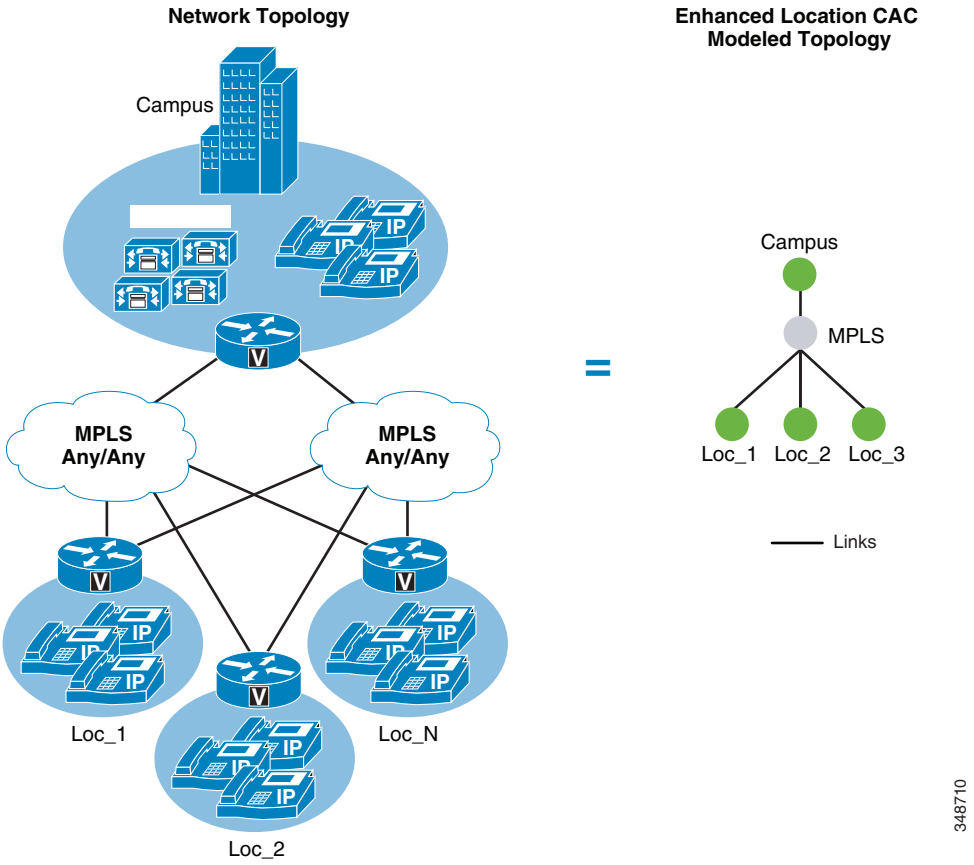


Figure 13-55 shows multiple MPLS clouds from different providers, where each site has one connection to each cloud and uses the MPLS clouds in either an equal-cost load-balanced manner or in a primary/backup scenario. In any case, this design is equivalent to the dual data center design where a single location represents both clouds and a single link represents the lowest capacity link of the two.

Figure 13-55 Remote Sites Connected to Dual MPLS Clouds



Design Recommendations

- The MPLS cloud should be configured as a location that does not contain any endpoints but is used as a hub to interconnect locations.
- The MPLS cloud serves as a transit location for interconnecting multiple variable-sized bandwidth WAN uplinks to other remote locations.
- Remote sites with connectivity to dual MPLS clouds should treat those connections as a single link and size to the lowest capacity of the links in order to avoid oversubscription during network failure conditions.

Call Admission Control Design Recommendations for Video Deployments

This section discusses Enhanced Location CAC and the design considerations and recommendations applicable to Quality of Service (QoS) when designing video deployments.

Admission control and QoS are complementary and in most cases co-dependent. Current Cisco product offerings such as audio and video endpoints, voice and video gateways, voice messaging, and conferencing all support native QoS packet marking based on IP Differentiated Services Code Point (IP DSCP). Note, however, that Jabber for Windows clients specifically do not follow the same native marking ability that other clients do, due to how the Windows operating systems requires the use of

Group Policy Objects (GPO) using application, IP addresses, and UDP/TCP port ranges to mark traffic with DSCP. Group Policy Objects are very similar in function to network access lists in their ability to mark traffic.

QoS is critical to admission control because without it the network has no way of prioritizing the media to ensure that admitted traffic gets the network resources that it requires above that of non-admitted or other traffic classifications. In Unified CM's CallManager service parameters for QoS, there are five main QoS settings that are applicable to endpoint media classification and that also allow immersive and desktop classified endpoints (see the section on Enhanced Location CAC for TelePresence Immersive Video, page 13-59) to have different QoS markings for their media based on their video classification of immersive or desktop. Table 13-14 shows the five main DSCP settings along with their default settings and Per Hop Behavior (PHB) equivalents.

Table 13-14 QoS settings for Endpoint Media Classification

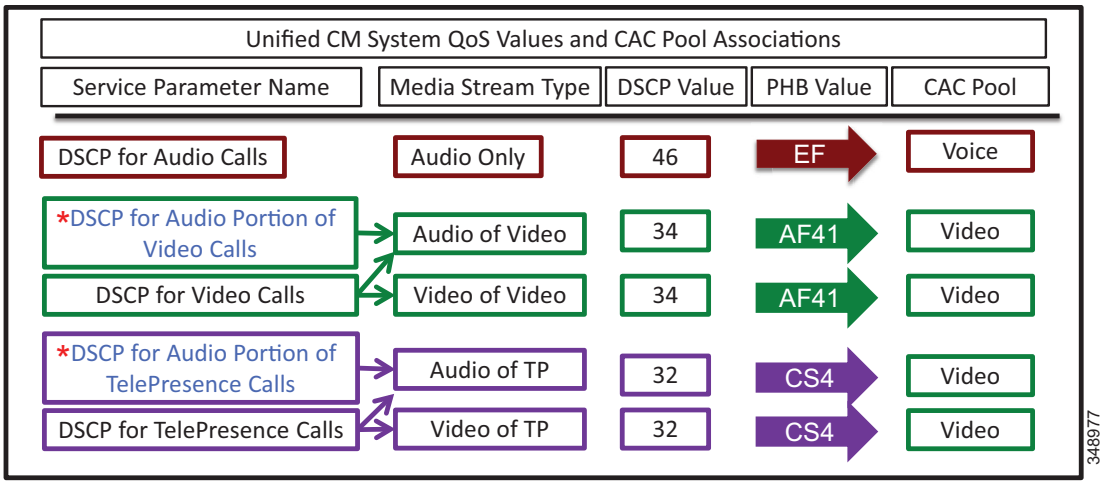
Cisco CallManager Service parameters > Clusterwide Parameters (System - QOS)	Default Value	PHB Equivalent
DSCP for Audio Calls	46	EF
DSCP for Video Calls	34	AF41
DSCP for Audio Portion of Video Calls	34	AF41
DSCP for TelePresence Calls	32	CS4
DSCP for Audio Portion of TelePresence Calls	32	CS4

The **DSCP for Audio Calls** setting is used for any device that makes an audio-only call. The **DSCP for Video Calls** setting is used for the audio and video traffic of any device that is classified as "desktop." **DSCP for TelePresence Calls** is used for the audio and video traffic of any device that is classified as "immersive." The **DSCP for Audio Portion of Video Calls** and **DSCP for Audio Portion of TelePresence Calls** are currently applicable to a subset of video endpoints and differentiate only the audio portion of video calls dependent on the video call type based on classification. See the section on Trusted Endpoints, page 13-93, for more information.

As mentioned in the section on Enhanced Location CAC for TelePresence Immersive Video, page 13-59, Cisco Unified CM E-LCAC has the ability to perform admission control for TelePresence calls separately from other video calls. E-LCAC does this through a classification of endpoints and SIP trunks as "immersive" or "desktop." This classification gives Unified CM the ability to deduct bandwidth from a separate immersive bandwidth pool for those devices and trunks classified as immersive. By default LBM deducts ALL video, no matter the classification, from the video bandwidth pool (Unified CM's CallManager service parameter **Use Video BandwidthPool for Immersive Video Calls** set to **True**).

Also by default, all immersive classified endpoints have a DSCP set to CS4 (DSCP 32; **DSCP for TelePresence Calls**), while desktop endpoints have a DSCP set to AF41 (DSCP 34; **DSCP for Video Calls**). The default settings for QoS and E-LCAC differentiate DSCP but deduct all video from the same E-LCAC bandwidth pool. Figure 13-56 illustrates the QoS and E-LCAC bandwidth pool associations and defaults for immersive and desktop classified devices.

Figure 13-56 Default QoS Settings for CAC Bandwidth Pools



DSCP for TelePresence Calls is the immersive classification, and **DSCP for Video Calls** is the desktop classification.

Enhanced Location CAC Design Considerations and Recommendations

When designing Enhanced Location CAC for video, follow the design recommendations and considerations listed in this section.

Design Recommendations

The following design recommendations apply to video solutions that employ Enhanced Location CAC:

- If you are deploying Unified Communications video (desktop classification) and TelePresence video (immersive classification) where differentiation between desktop video and TelePresence video is a requirement, then ensure that the Unified CM service parameter **Use Video Bandwidth Pool for Immersive Video Calls** is set to **false**. This enables the immersive bandwidth pool for TelePresence calls.
- In Enhanced Location CAC, TelePresence endpoints can be managed in the same location as Unified Communications video endpoints. If TelePresence calls are not to be tracked through Enhanced Location CAC, then set the immersive location and links bandwidth pool to **unlimited**. This will ensure that CAC will not be performed on TelePresence or SIP trunks classified as immersive. If TelePresence calls are to be tracked through Enhanced Location CAC, then set immersive location and links bandwidth pool to a value according to the bit rate used and the number of calls to be allowed over the locations and link paths.
- Intercluster SIP trunks should be associated with the shadow location.
- Cisco Unified CM uses two different cluster-wide QoS service parameter to differentiate between the Differentiated Services Code Point (DSCP) settings of UC video endpoints and TelePresence endpoints. TelePresence endpoints use the **DSCP for Telepresence calls** QoS parameter while the Cisco UC video endpoints use the **DSCP for video calls** QoS service parameter.

- When marking video with the default QoS markings, the following recommendations apply:
  - For sites that deploy only UC endpoints and no TelePresence endpoints, ensure that the CS4 DSCP class is added to the AF41 QoS traffic class on inbound WAN QoS configurations to account for the inbound CS4 marked traffic, thus ensuring QoS treatment of CS4 marked media.
  - For sites that deploy only UC TelePresence endpoints and no UC endpoints, ensure that the AF41 DSCP class is added to the CS4 QoS traffic class on inbound WAN QoS configurations to account for the inbound AF41 marked traffic, thus ensuring QoS treatment of AF41 marked media.

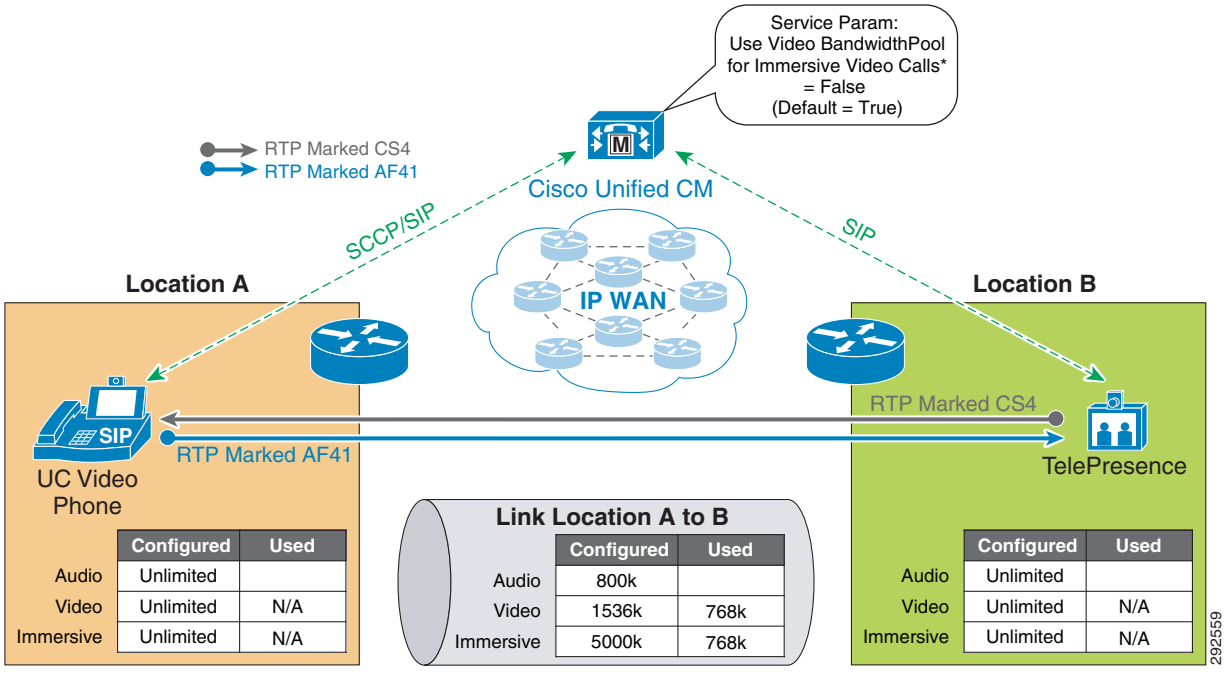
Design Considerations

When deploying Enhanced Location CAC for immersive video calls, consider the effects of DSCP marking for both QoS classes, as the interoperable calls where an immersive classified endpoint is connected with a desktop classified endpoint are by default asymmetrically marked.

DSCP QoS Marking

The Differentiated Services Code Point (DSCP) QoS markings for TelePresence video interoperable calls are asymmetric, with AF41 used for the UC endpoints and CS4 for the TelePresence endpoints. AF41 and CS4 are default configurations in Unified CM, and changes to these defaults should align with the QoS configuration in the network infrastructure, as applicable. TelePresence endpoints mark video calls with a DSCP value of CS4, which is consistent with the default **DSCP for Telepresence calls** setting. UC endpoints mark calls with a DSCP value of AF41, which is consistent with the default **DSCP for Video calls** setting. Figure 13-57 illustrates the media marking and bandwidth accounting.

Figure 13-57 Bandwidth Deductions and Media Marking in a Multi-Site Deployment with Enhanced Location CAC



Bandwidth Accounting for TelePresence Video Interoperability Calls

Enhanced Location CAC for TelePresence-to-UC video interoperable calls deducts bandwidth from both the video and immersive locations and links bandwidth pools, as illustrated in Figure 13-57. This is by design to ensure that both types of QoS classified streams have the bandwidth required for media in both directions of the path between endpoints.

Enhanced Location CAC accounts for the bidirectional media of both AF41 and CS4 class traffic. In asymmetrically marked flows, however, the full allocated bit rate of the AF41 class is used in one direction but not the other. In the other direction, the full allocated bit rate is marked CS4. This does not represent additional bandwidth consumption but simply a difference in marking and queuing in the network for each QoS class. This manner of bandwidth accounting is required to protect each flow in each direction.

If TelePresence video (CS4) has been provisioned in the network paths separately from Unified Communications video (AF41) and TelePresence is largely scheduled and in environments where the scheduling of calls is controlled and the utilization of TelePresence is deterministic, then immersive video bandwidth for locations and links can be set to **unlimited** to avoid the double bandwidth CAC calculations. This ensures that TelePresence-to-TelePresence calls always go through unimpeded and will not be subject to admission control, while desktop video and TelePresence-to-desktop video calls will be subject to admission control and accounted for in the video bandwidth allocation.

For more information on the call flows for Enhanced Location CAC and TelePresence interoperable calls, see the section on Enhanced Location CAC for TelePresence Immersive Video, page 13-59.

Design Recommendations for Unified CM Session Management Edition Deployments with Enhanced Location CAC

Unified CM Session Management Edition (SME) is typically used for interconnecting multiple Unified CM clusters, third-party UC systems (IP- and TDM-based PBXs), PSTN connections, and centralized UC applications as well as for dial-plan and trunk aggregation. The following is a list of recommendations and design considerations to follow when deploying Unified CM SME with Enhanced Location CAC. For more information on Unified CM SME, see the chapter on Collaboration Deployment Models, page 10-1.

Recommendations and Design Considerations

- All leaf clusters that support Enhanced Location CAC should be enabled for intercluster Enhanced Location CAC with SME.
- SME can be used as a centralized bootstrap hub for the Enhanced Location CAC intercluster hub replication network. See LBM Hub Replication Network, page 13-52, for more information.
- All trunks to leaf clusters supporting Enhanced Location CAC should be SIP trunks placed in the shadow location to enable Enhanced Location CAC on the trunk between SME and the leaf clusters supporting Enhanced Location CAC.
- For TelePresence video interoperability, see the section on Call Admission Control Design Recommendations for Video Deployments, page 13-78.
- Connectivity from SME to any trunk or device other than a Unified CM that supports Enhanced Location CAC (some examples are third-party PBXs, gateways, Unified CM clusters that support only traditional Location CAC, voice messaging ports or trunks to conference bridges, Cisco Video Communications Server, and so forth) should be configured in a location other than a phantom or shadow location. The reason for this is that both phantom and shadow locations are non-terminating

locations; that is, they relay information about locations and are effectively placeholders for user-defined locations on other clusters. Phantom locations are legacy locations that allow for the transmission of location information in versions of Unified CM that support only traditional Location CAC, but they are not supported with Unified CM Enhanced Location CAC. Shadow locations are special locations that enable trunks between Unified CM clusters that support Enhanced Location CAC to accomplish it end-to-end.

- SME can be used as a locations and link management cluster. See Figure 13-58 as an example of this.
- SME can support a maximum of 2,000 locations configured locally.

Figure 13-58 Unified CM SME as a Location and Link Management Cluster

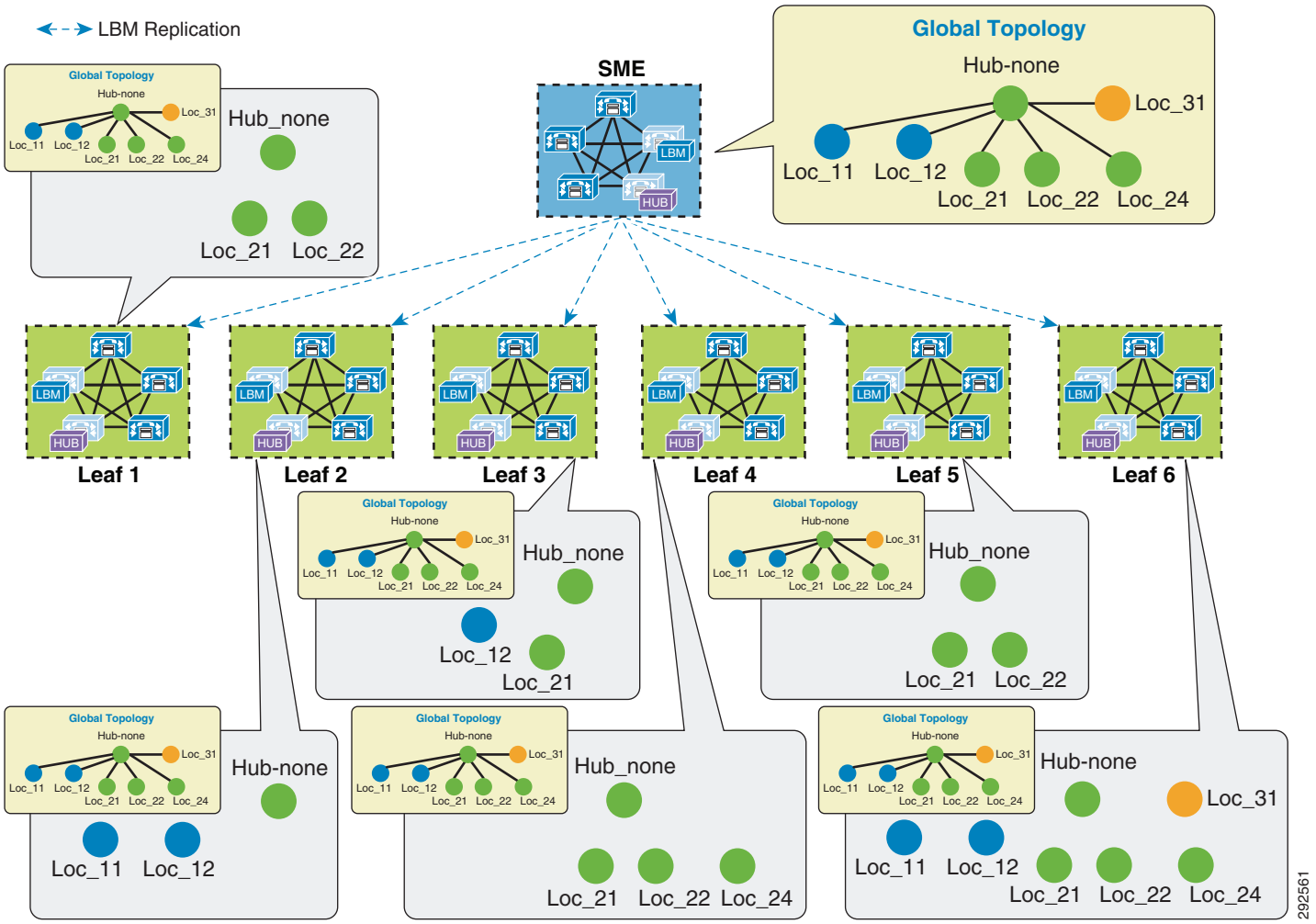
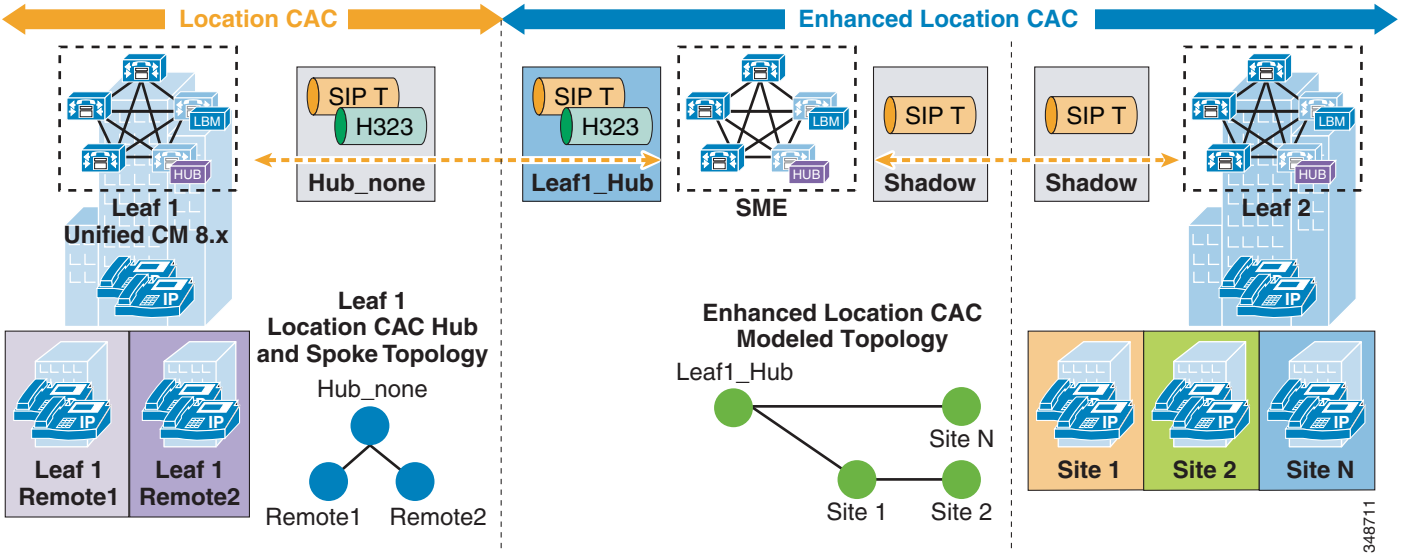


Figure 13-58 illustrates SME as a location and link management cluster. The entire location and link global topology is configured and managed in SME, and the leaf clusters configure locally only the locations that they require to associate with the end devices. When intercluster Enhanced Location CAC is enabled and locations and links are replicated, each leaf cluster will receive the global topology from SME and overlay this on their configured topology and use the global topology for call admission

control. This simplifies configuration and location and link management across multiple clusters, and it diminishes the potential for misconfiguration across clusters. For more information and details on the design and deployment see the section on Location and Link Management Cluster, page 13-56.

Figure 13-59 illustrates an SME design where intercluster Enhanced Location CAC has been enabled on one or more leaf clusters (right) and where one or more leaf clusters are running a version of Unified CM that supports only traditional Location CAC (left). In this type of a deployment the locations managed by traditional Location CAC cannot be common or shared locations between clusters enabled for Enhanced Location CAC. Leaf 1 has been configured in a traditional hub and spoke, where devices are managed at various remote sites. SME and the other leaf clusters that are enabled for intercluster Enhanced Location CAC share a global topology, as illustrated in the E-L CAC Modeled Topology. Leaf1_Hub is a user-defined location in SME assigned to the SIP or H.323 intercluster trunk that represents the hub of the Leaf 1 topology. This allows SME to deduct bandwidth for calls to and from Leaf 1 up to the Leaf1_Hub. In this way SME and Leaf 2 manage the Enhanced Location CAC locations and links while Leaf 1 manages its remote locations with traditional Location CAC.

Figure 13-59 SME Design with Enhanced Location CAC and Traditional Location CAC in Leaf Clusters



Design Recommendations for Cisco Expressway Deployments with Enhanced Location CAC

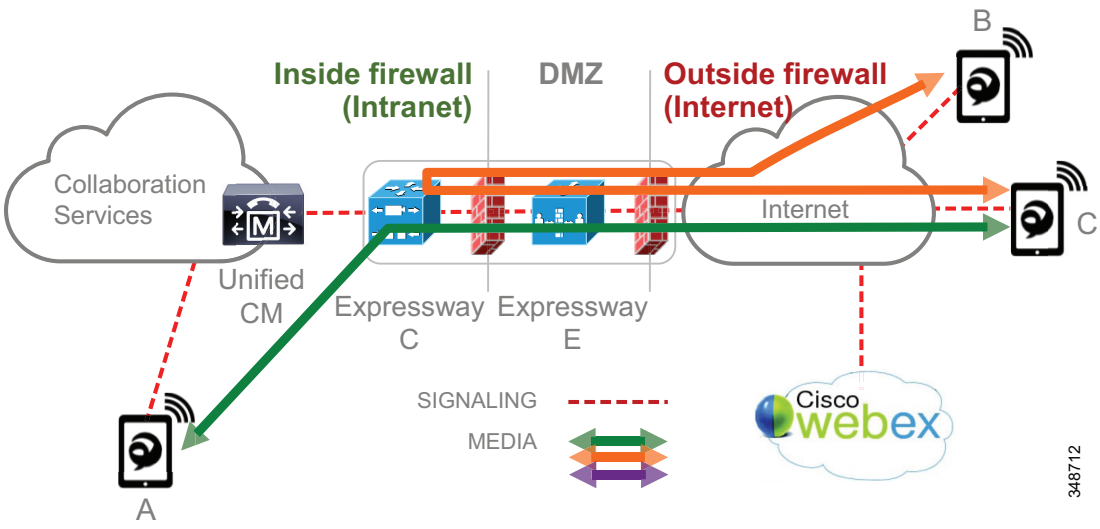
Cisco Expressway mobile and remote access capabilities provide registration of Internet-based devices to Unified CM without the use of a VPN, otherwise known as VPN-less enterprise access. This allows the endpoint or client application to register securely to Unified CM without the need for the entire operating system hosting the application to have access to the enterprise network. The following section lists the recommendations and design considerations for deploying mobile and remote access with Enhanced Location Call Admission Control (ELCAC). For more information on mobile and remote access, refer to the section on VPN-less Enterprise Access, page 10-36.

Recommendations and Design Considerations

In the Cisco Expressway VPN-less mobile and remote access solution, endpoints supporting the feature can register to Unified CM through a Cisco Expressway deployment without the use of a VPN. Cisco Expressway C and Expressway E servers are deployed, each with redundancy for high availability. Expressway E is placed in the DMZ between the firewall to the Internet (outside) and the firewall to the enterprise (inside), while Expressway C is placed inside the enterprise. Figure 13-60 illustrates this deployment. It also illustrates the following media flows:

1. For Internet-based endpoints calling one another, the media is routed through Cisco Expressway E and Expressway C back out to the Internet, as is illustrated between endpoints B and C in Figure 13-60.
2. For Internet-based endpoints calling internal endpoints, the media flows through the Expressway E and Expressway C, as is illustrated between endpoints A and C in Figure 13-60.

Figure 13-60 Deployment of Cisco Expressway for VPN-less Access



For multiple deployments of Cisco Expressway for VPN-less access in the same enterprise, with the Internet-based endpoints registered through one Expressway pair calling Internet-based endpoints registered through another Expressway pair, the media will be routed through the enterprise. This is illustrated in Figure 13-61 with a call between endpoint D and endpoint C, both registered from the Internet but through two different Expressway pairs. The media flow will be the same whether the endpoints are registered to the same Unified CM cluster or to different Unified CM clusters.

Figure 13-61 Media Flow for a Deployment of Multiple Cisco Expressway Pairs

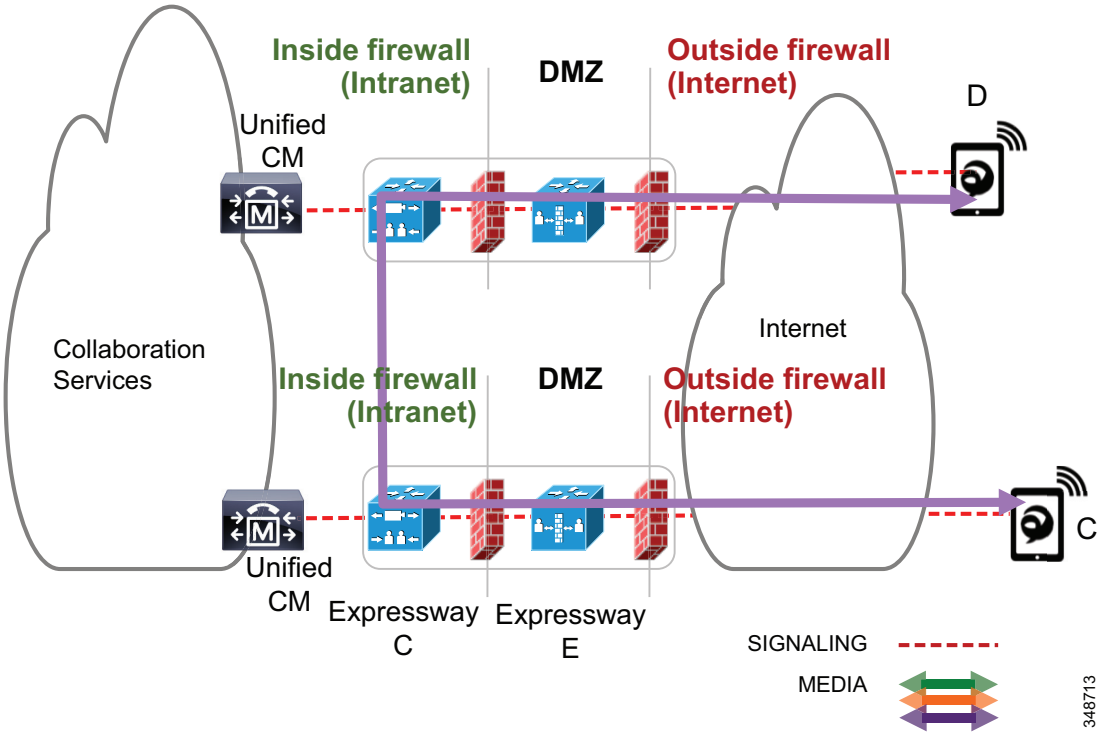


Figure 13-62 illustrates an example configuration for locations and links that integrate bandwidth tracking for media flows that traverse the enterprise, while still allowing media flows over the Internet without admission control.

Figure 13-62 Locations and Links for Remote and Mobile Access

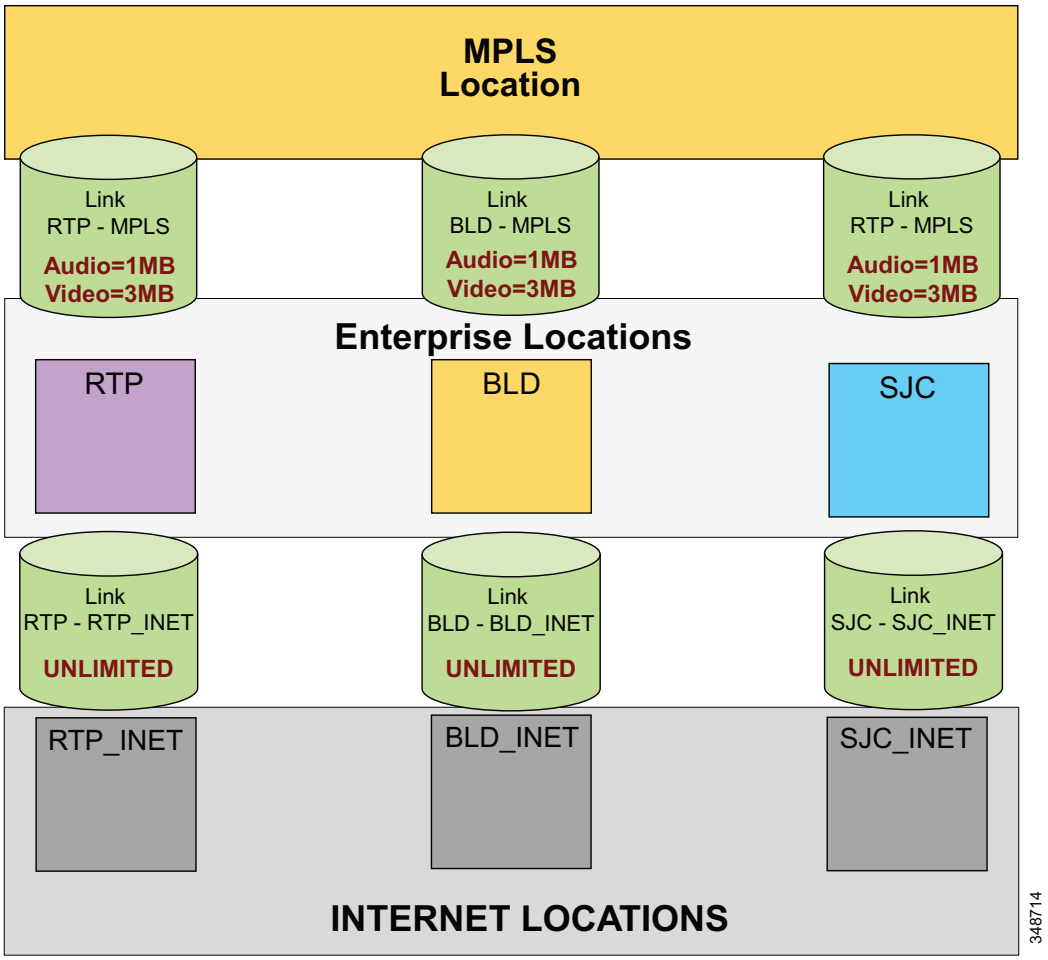


Figure 13-62 illustrates an example deployment of ELCAC consisting of three main sites: RTP, BLD, and SJC. These sites are all connected to an MPLS provider and thus each has a separate WAN connection to the MPLS cloud. Locations and links are created accordingly so that the enterprise locations are linked directly to a location called MPLS, with bandwidth links limited for audio and video calls mapping to the network topology. Devices are located in one of the three sites when in the enterprise and thus have a location associated to them. Each of these sites has a Cisco Expressway solution for VPN-less remote and mobile access for Internet-based endpoints registering to Unified CM. Three new locations are configured for the Internet-based devices, one for each Expressway solution site, named RTP_INET, BLD_INET, and SJC_INET. These three locations represent "Internet locations" because they are locations for devices registering from the Internet to Unified CM through an Expressway pair. These locations are not interconnected with direct links. This is because calls between Expressways are routed through the enterprise and thus flow through the MPLS cloud. These Internet locations, instead, have a link to their associated enterprise location. For example, RTP_INET has a link to RTP, BLD_INET has a link to BLD, and so forth. These links between the Internet locations and the enterprise locations should be set to **unlimited** bandwidth.

As mentioned, Enhanced Location CAC for Cisco Expressway deployments requires the use of a feature in Unified CM called Device Mobility. (For details about this feature, see the section on Device Mobility, page 21-14.) Enabling device mobility on the endpoints allows Unified CM to know when the device is

registered through the Cisco Expressway or when it is registered from within the enterprise. Device mobility also enables Unified CM to provide admission control for the device as it roams between the enterprise and the Internet. Device mobility is able to do this by knowing that, when the endpoints register to Unified CM with the IP address of Expressway C, Unified CM will associate the applicable Internet location. However, when the endpoint is registered with any other IP address, Unified CM will use the enterprise location that is configured directly on the device (or from the device pool directly configured on the device). It is important to note that device mobility does not have to be deployed across the entire enterprise for this function to work. Configuration of Device Mobility in Unified CM is required only for the Expressway IP addresses, and the feature is enabled only on the devices that require the function (that is to say, those devices registering through the Internet). Figure 13-63 illustrates an overview of the device mobility configuration. Although this is a minimum configuration requirement for Device Mobility for ELCAC to function for internet-based devices, Device Mobility can be configured to support mobility for these same endpoints within the enterprise. (See the section on Device Mobility, page 21-14, for more information.)

Figure 13-63 Device Mobility Configuration and Location Association

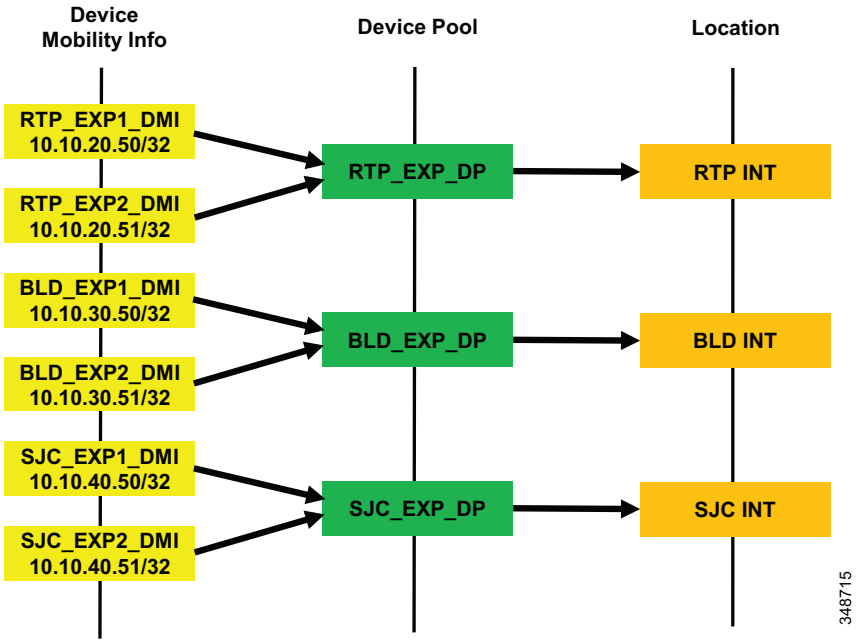
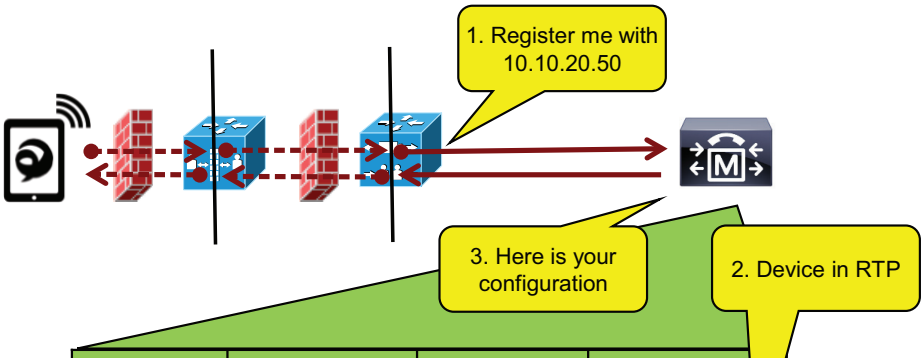


Figure 13-63 shows a simplified version of device mobility for the example deployment of ELCAC as described in Figure 13-62. The IP addresses of the Expressway C servers are configured in the device mobility information. In this example there is a redundant pair of Expressway C servers for each of the three sites, RTP, BLD, and SJC. RTP_EXP1_DMI and RTP_EXP2_DMI are configured respectively with the server IP addresses of the RTP Expressway C servers. These two are associated to a new device pool called RTP_EXP_DP, which has the location RTP_INET configured on it. Each site is configured similarly. With this configuration, when any device enabled for device mobility registers to Unified CM with the IP Address that corresponds to the device mobility information in RTP_EXP1_DMI or RTP_EXP2_DMI, it will be associated with the RTP_EXP_DP device pool and thus with the RTP_INET location.

With the above configuration, when an Internet-based device registers through the Expressway to Unified CM, it will register with the IP address of Expressway C. Unified CM then uses the IP address configured in the device mobility information and associates the device pool and thus the Internet location associated to this device pool. This process is illustrated in Figure 13-64.

Figure 13-64 Association of Device Pool and Location Based on Expressway IP Address



IP Subnet	Device Mobility Info	Device Pool	Location
10.10.20.50	RTP_EXP1_DMI	RTP_EXP_DP	RTP_INET
10.10.30.50	BLD_EXP1_DMI	BLD_EXP_DP	BLD_INET
10.10.40.50	SJC_EXP1_DMI	SJC_EXP_DP	SJC_INET

348716

In Figure 13-64 the client registers with Unified CM through the Expressway in RTP. Because the signaling is translated at the Expressway C in RTP, the device registers with the IP address of the Expressway C. The device pool RTP_EXP_DP is associated to the device based on this IP address. The RTP_EXP_DP pool is configured with the RTP_INET location, and therefore that location is associated to the device. Thus, when devices register to the Expressway, they get the correct location association through device mobility. When the endpoint relocates to the enterprise, it will return to its static location configuration. Also, if the endpoint relocates to another Expressway in SJC, for example, it will get the correct location association through device mobility.

Design and Deployment Best Practices for Cisco Expressway VPN-less Access with Enhanced Location CAC

- Each site with Internet access, where a Cisco Expressway solution resides, requires an Internet location and an enterprise location. Each Cisco Expressway deployment requires these location pairs. The enterprise location is associated to devices when they are in the enterprise (see locations RTP, BLD, and SJC in Figure 13-62). The Internet location is associated to the endpoints through the Device Mobility feature when the endpoints are registering from the Internet (see locations RTP_INET, BLD_INET, and SJC_INET in Figure 13-62). For example, in Figure 13-62, RTP and RTP_INET form a location pair for the physical site RTP.
- Enterprise locations are configured according to applicable enterprise ELCAC design.
- Internet locations will always have a single link to the enterprise location that they are paired with. For example, in Figure 13-62, RTP and RTP_INET form an enterprise location and internet location pair.
- Links from Internet locations to enterprise locations are set to **unlimited** bandwidth. Unlimited bandwidth between these location pairs ensures that bandwidth is not counted for calls from the Internet location to the local enterprise location, and vice versa (for example, calls from RTP to RTP_INET in Figure 13-62).
- In a Cisco Expressway solution where more than one Cisco Expressway site is deployed, and requiring multiple Internet locations, ensure that Internet locations do not have direct links between one another. Direct links between Internet locations will create multiple paths in ELCAC, and for that reason they are not recommended.
- When configuring DSCP on Cisco Expressway, ensure that it is consistent with your endpoint marking policy. Starting with Cisco Expressway Release 8.9, DSCP can be configured separately for signaling, audio, video, and XMPP. Thus, it is possible to configure signaling as CS3 (24), audio as EF (46), video as AF41 (32), and XMPP as CS3 (24), and the Expressway-C will appropriately mark the media and signaling traffic coming from the Internet into the Enterprise. When you upgrade an Expressway server from a release prior to 8.9, the single configured DSCP value used in the earlier release will be populated across all 4 values in the new Expressway release.

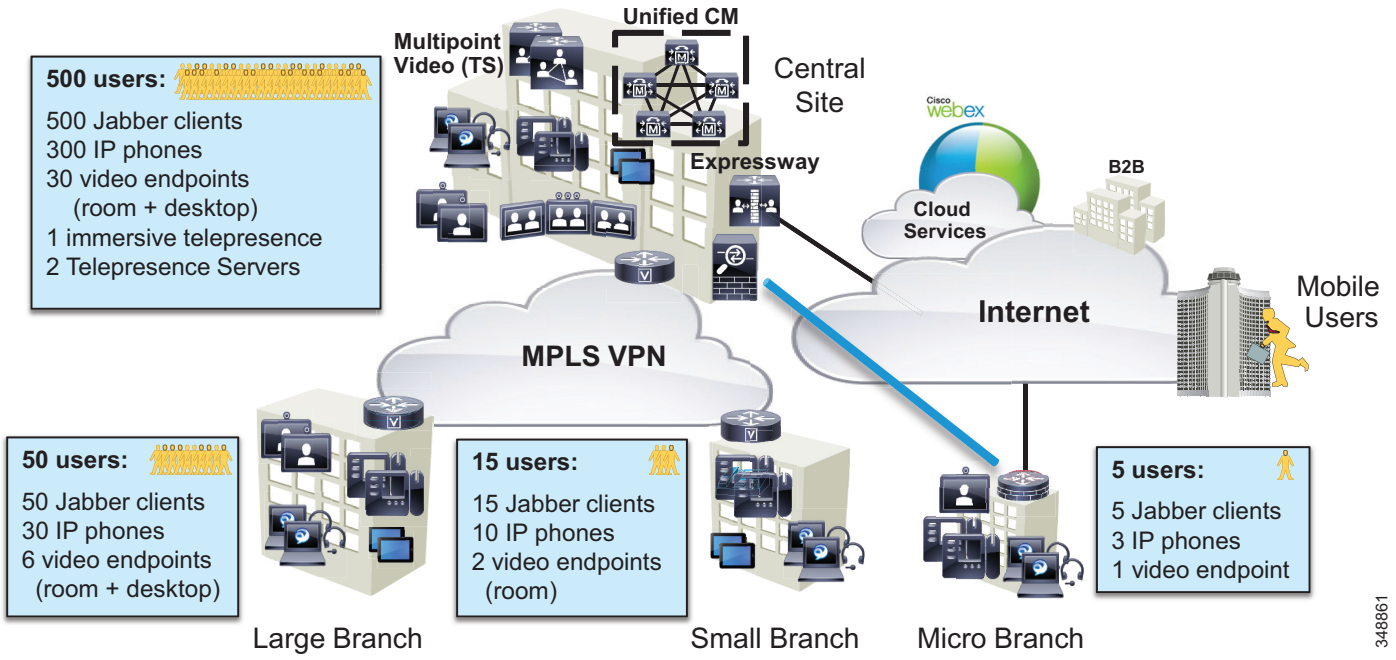
Bandwidth Management Design Examples

This section covers design examples and explores all aspects discussed in this chapter – identification and classification, WAN queuing and scheduling, provisioning and resource control, and bandwidth allocation guidelines – with details for each site in the examples.

Example Enterprise #1

Example Enterprise #1 is a large enterprise with users across a large geographic area, with a data center (DC) at the headquarters site as well as multiple large, small, and micro-sized branches with roughly 500, 50, 15, and 5 users in each branch type, respectively. To simplify the illustration of the network, these categories of sites (HQ, large, small and micro) are used as a template to size bandwidth considerations for each site that has a similar user base and endpoint density. Figure 13-65 illustrates each type of site. The enterprise has deployed Jabber with video to ensure that users have access to a video terminal for conferencing. The TelePresence video conferencing resources are located in the DC at HQ. IP phones are for voice-only communications; video endpoints are Jabber clients, Collaboration desktop endpoints (DX Series), and room endpoints (MX, Profile, and SX Series); and the HQ and large sites have immersive TelePresence units such as the IX Series.

Figure 13-65 Example Enterprise #1



The IT department is tasked with determining the bandwidth requirements for the WAN edge for each type of site in Example Enterprise #1. The following sections list the requirements and illustrate a methodology for applying QoS and for determining bandwidth and queuing requirements as well as admission control requirements.

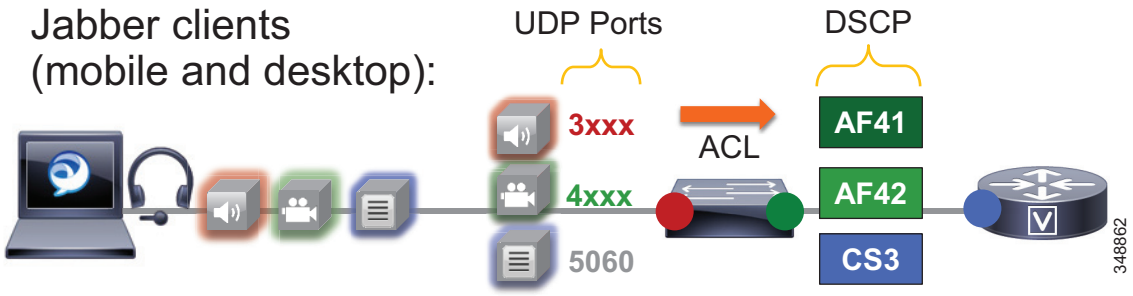
Identification and Classification

In this phase the QoS requirements are established across the enterprise.

Untrusted Endpoints (Jabber)

Jabber endpoints are untrusted and sit in the data VLAN. Specific UDP port ranges will be used to re-mark signaling and media at the access layer switch. In this case Unified CM is configured with a SIP Profile specifically for all Jabber clients to use the Common Media and Signaling Port Range of 3000 to 4999. This sets all Jabber endpoints to use a source UDP port of 3000 to 3999 for audio streams and 4000 to 4999 for video streams. The default SIP port of 5060 is used for SIP signaling (configured in the SIP Security Profile). This is illustrated in Figure 13-66.

Figure 13-66 Untrusted (Jabber) Endpoint QoS



The administrator creates an ACL for the access switches for the data VLAN to re-mark UDP ports to the following DSCP values:

- Audio: UDP Ports 3000 to 3999 marked to AF41
- Video: UDP Ports 4000 to 4999 marked to AF42
- Signaling: TCP Port 5060 marked to CS3

Jabber classification summary:

- Audio streams of all Jabber calls (voice-only and video calls) are marked AF41.
- Video streams of Jabber video calls are marked AF42.

For the Jabber endpoints, we also recommend changing the default QoS values in the Jabber SIP profile. This is to ensure that, if for any reason the QoS of a Jabber client is trusted via a wireless route or any other wired route, the correct trusted values will be coherent between the trusted QoS and the QoS that is re-marked with the ACLs. Therefore, the QoS parameters in the SIP Profile for Jabber clients need to be set as shown in Table 13-15.

Table 13-15 QoS Parameters in SIP Profile for Jabber Clients

QoS Service Parameter Name (SIP Profile)	System Default Value	Changed Value
DSCP for Audio Calls	EF	AF41
DSCP for Video Calls	AF41	AF42
DSCP for Audio Portion of Video Calls	AF41	

Table 13-15 QoS Parameters in SIP Profile for Jabber Clients (continued)

QoS Service Parameter Name (SIP Profile)	System Default Value	Changed Value
DSCP for TelePresence Calls	CS4	N/A
DSCP for Audio Portion of TelePresence Calls	CS4	N/A

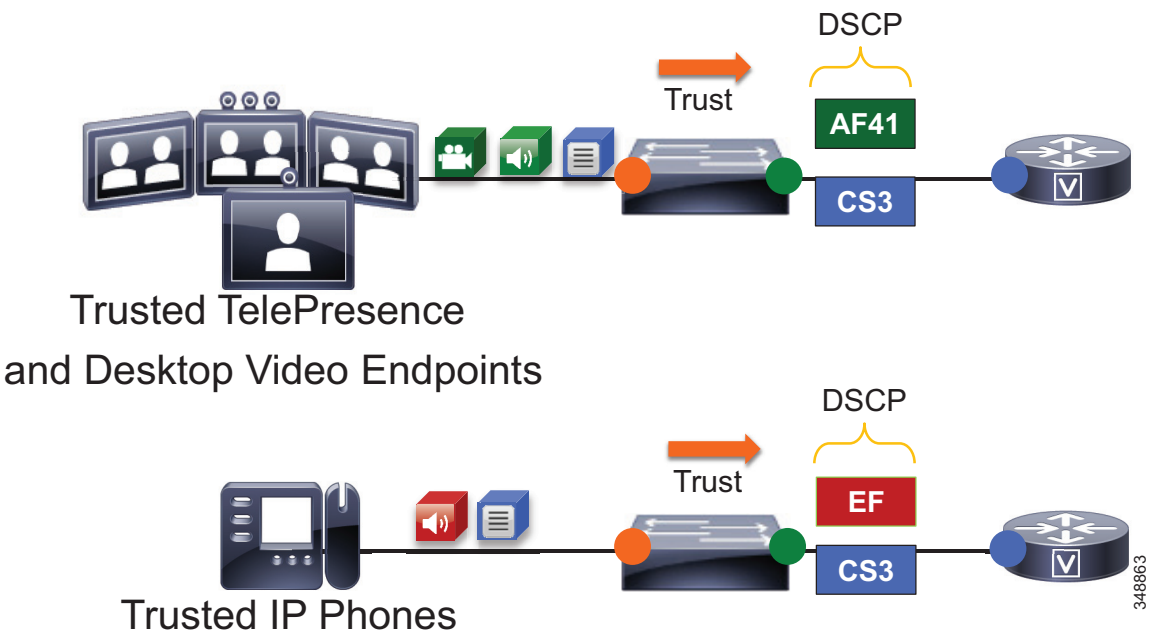
The configuration settings in Table 13-15 ensure that video of Jabber clients will be set to AF42 if for any reason the traffic follows a trusted network path and is not re-marked via UDP port ranges as in the untrusted network path. The DSCP for Audio Portion of Video Calls is left at the default setting of AF41. This is simply to ensure a consistent configuration across Jabber endpoints, whether trusted or re-marked via the network using UDP port ranges.

Trusted Endpoints

For the trusted endpoints, Cisco Discovery Protocol (CDP) is used and the QoS of the IP phones and video endpoints is trusted using the conditional trust mechanism configured at the access switch. The configuration uses the Unified CM default system settings of audio for voice-only calls as EF, audio and video for video calls as AF41, audio and video for TelePresence as CS4, and signaling as CS3. Therefore, the administrator must change the QoS defaults in Unified CM for the trusted endpoints with a SIP Profile to ensure that the QoS of the TelePresence endpoints is adjusted accordingly.

Figure 13-67 illustrates the conditional trust (CDP based) and packet marking at the access switch.

Figure 13-67 Trusted Endpoint QoS



The administrator configures all access switches with a conditional QoS trust for IP phones and video and TelePresence endpoints, classified as follows:

- Audio and video streams of video calls are marked AF41.
- Voice-only calls are marked EF.

The administrator must also change the QoS defaults in Unified CM for the trusted endpoints with a SIP Profile using the values in Table 13-16.

Table 13-16 QoS Parameters in SIP Profile for Trusted Endpoints

QoS Service Parameter Name (SIP Profile)	System Default Value	Changed Value
DSCP for Audio Calls	EF	
DSCP for Video Calls	AF41	
DSCP for Audio Portion of Video Calls	AF41	
DSCP for TelePresence Calls	CS4	AF41
DSCP for Audio Portion of TelePresence Calls	CS4	AF41

On ingress at the WAN edge, it is expected that the packets arriving with a specific DSCP value have been trusted at the access layer or re-marked accordingly if they were not trusted at the access switch. As a failsafe practice, on ingress it is important to re-mark any untrusted traffic at the WAN edge that could not be re-marked at the access layer. While QoS is important in the LAN, it is paramount in the WAN; and as routers assume a trust on ingress traffic, it is important to configure the correct QoS policy that aligns with the business requirements and user experience. The WAN edge re-marking is always done on the ingress interface into the router, while the queuing and scheduling is done on the egress interface. The following example walks through the WAN ingress QoS policy as well as the egress queuing policy. Figure 13-68 illustrates the configuration and the re-marking process.

In Figure 13-68 the packets from both the trusted and untrusted areas of the network are identified and classified with the appropriate DSCP marking via the trust methods discussed or via a simple ACL matching on UDP port ranges. Keep in mind that this ACL could also match more granularly on IP addresses or some other attributes that would further limit the scope of the marking.

Figure 13-68 Example Router Ingress QoS Policy Process – Step 1

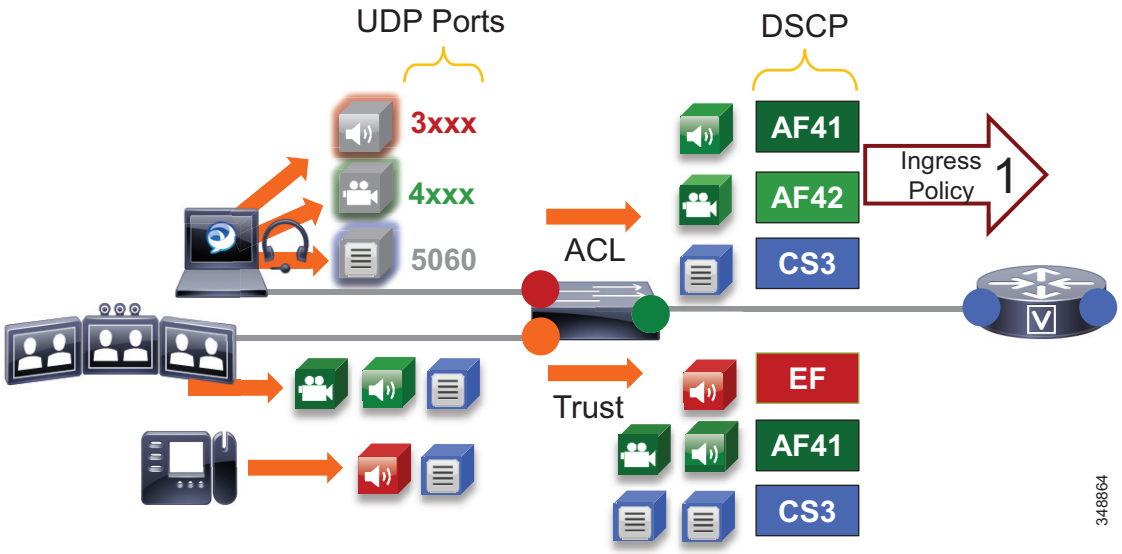
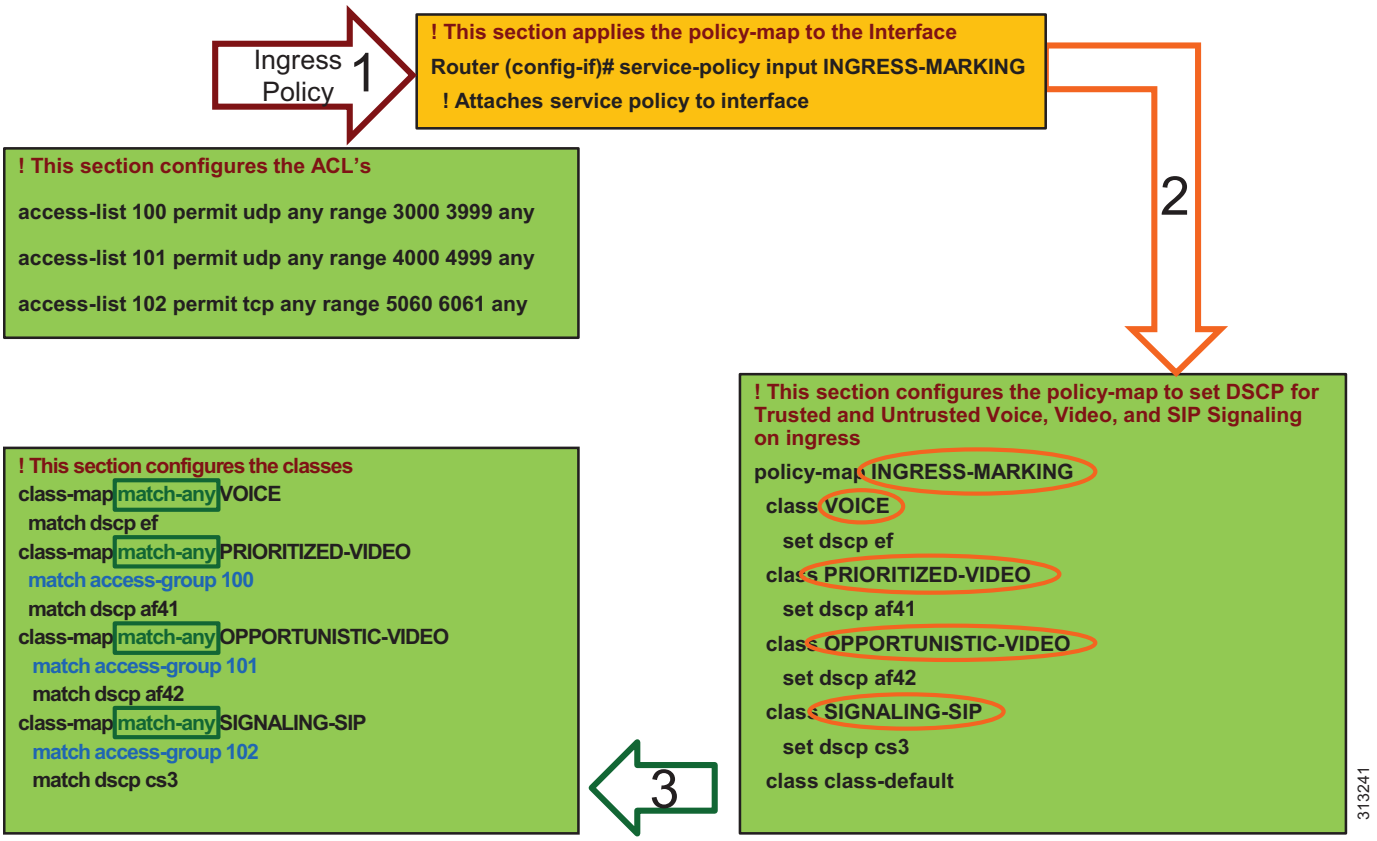


Figure 13-68 through Figure 13-73 illustrate the ingress QoS policy matching criteria and DSCP re-marking. The process involves the following steps shown in the figures:

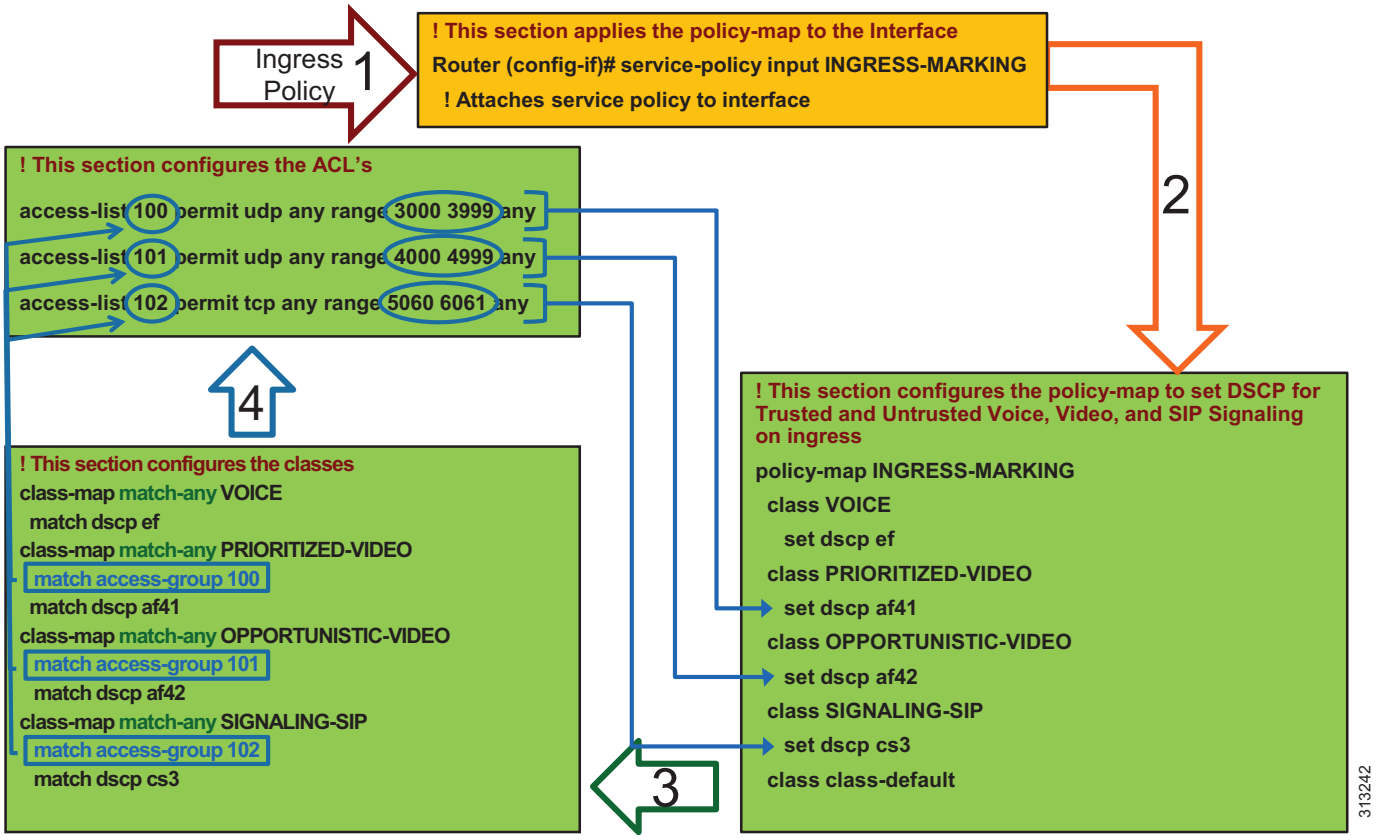
1. In step 1, packets arrive at the router ingress interface, which is configured with an input service policy (Figure 13-69).
2. In step 2, the policy-map is configured with four classes of traffic to set the appropriate DSCP: VOICE = EF; PRIORITIZED-VIDEO = AF41; OPPORTUNISTIC-VIDEO = AF42; SIGNALING-SIP = CS3 (Figure 13-69).
3. In step 3, each one of these classes matches a class-map of the same name configured with match-any criterion (Figure 13-69). This match-any criterion means that the process will start top-down, and the first matching criterion will be executed according to each class in the policy-map statements.

Figure 13-69 Example Router Ingress QoS Policy Process – Steps 1 to 3



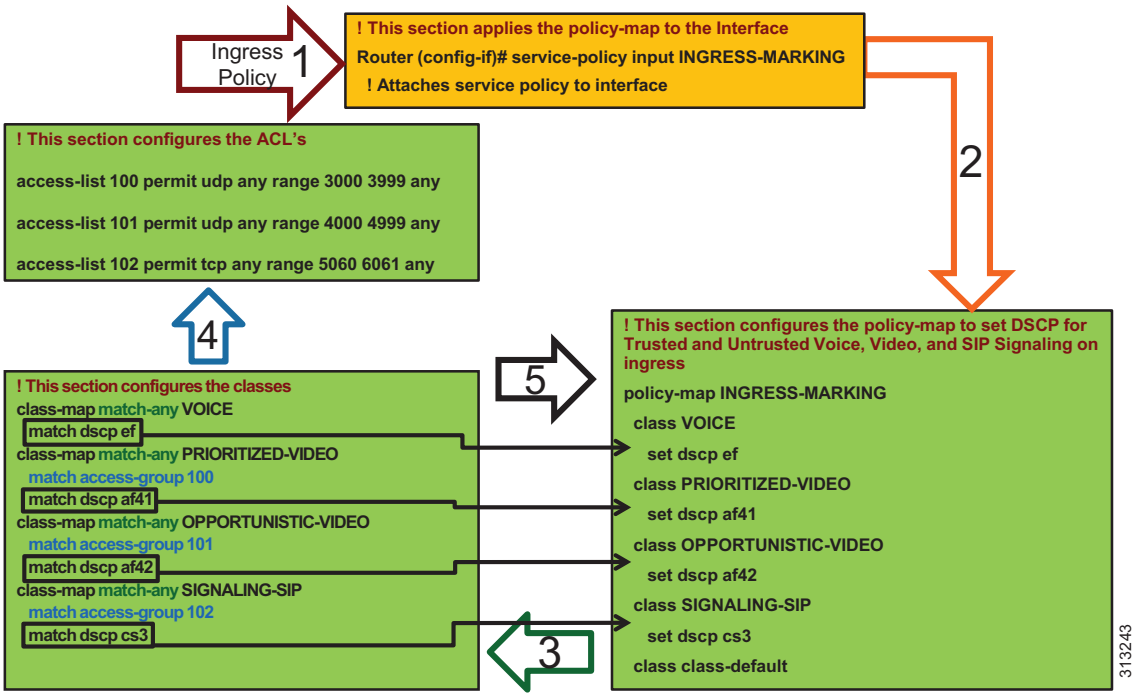
4. In step 4, the first line in the class-map statement is parsed, which is the ACL that matches the UDP ports set in Unified CM in the Identification and Classification section. When the ACL criteria are met (protocol and port range), then the traffic is marked as is configured in the corresponding policy-map statements (Figure 13-70). Note that Jabber Audio is marked AF41 and Jabber Video is marked AF42, in line with the policy in Figure 13-68.

Figure 13-70 Example Router Ingress QoS Policy Process – Step 4



5. In step 5, the traffic that did not match the first statement goes to the next match statement in the class-map, which is **match dscp** (Figure 13-71). If the traffic simply matches the DSCP, then DSCP is set again to the same value that was matched and as is configured in the policy-map statements. In this case the router is simply matching on DSCP and resetting the DSCP to the same value. This is a catch-all setting for the trusted DSCP from servers and applications coming into the WAN router.

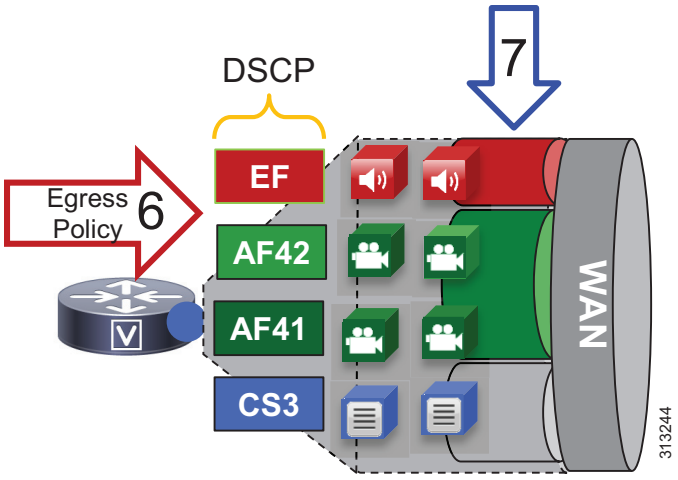
Figure 13-71 Example Router Ingress QoS Policy Process – Step 5



Note This is an example QoS ingress marking policy based on the Modular QoS CLI (MQC). Refer to your specific router configuration guide for information on how to achieve a similar policy on a Cisco router supporting MQC and for any updated commands.

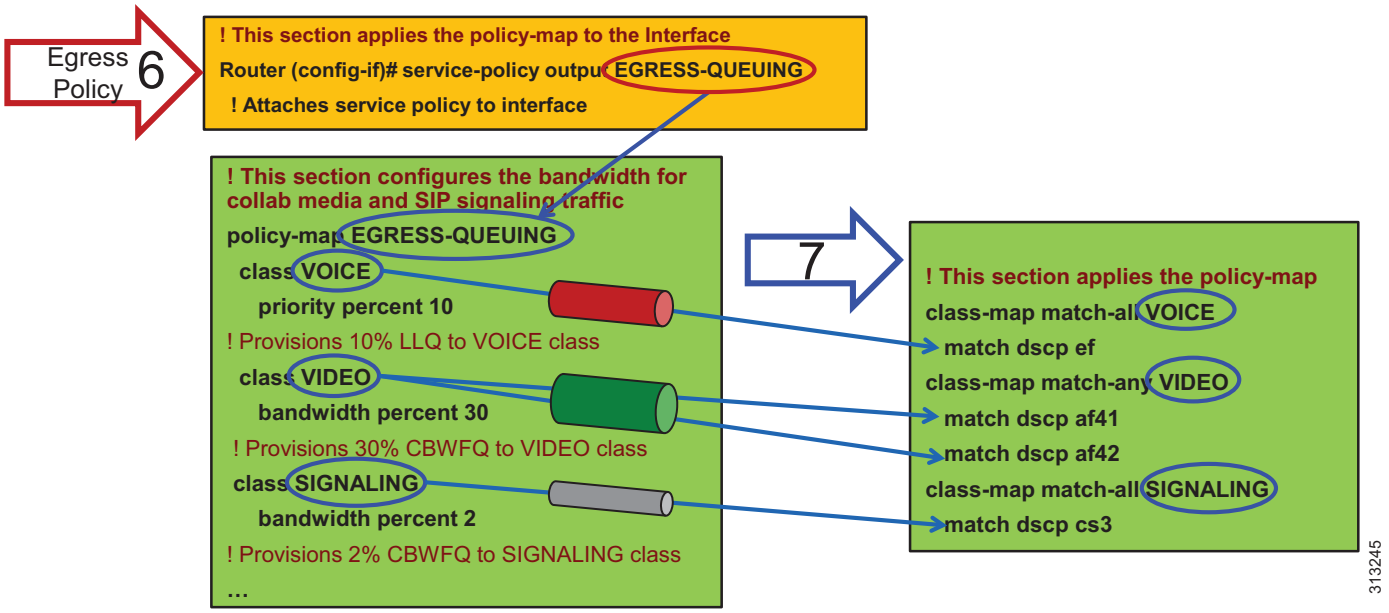
6. In step 6, the traffic goes to an outbound interface to be queued and scheduled by an output service policy that has three queues created: a Priority Queue called VOICE, a CBWFQ called VIDEO, and another CBWFQ called SIGNALING. This is illustrated in Figure 13-72 and Figure 13-73. This highlights the fact that the egress queuing policy is based only on DSCP as network marking occurring at the access switch and/or on ingress into the WAN router ingress interface. This is an example simply to illustrate the matching criteria and queues, and it does not contain the WRED functionality. For information on WRED, see the section on WAN Queuing and Scheduling, page 13-100.

Figure 13-72 Example Router Egress Queuing Policy Process – Step 6



7. In step 7, the traffic is matched against the class-map match statements (Figure 13-73). All traffic marked EF goes to the VOICE PQ, AF41 and AF42 traffic goes to the VIDEO CBWFQ, and CS3 traffic goes to the SIGNALING CBWFQ.

Figure 13-73 Example Router Egress Queuing Policy Process – Step 7



Note This is an example egress queuing policy based on the Cisco Common Classification Policy Language (C3PL). Refer to your specific router configuration guide for information on how to achieve a similar policy on a Cisco router supporting C3PL and for any updated commands.

WAN Queuing and Scheduling

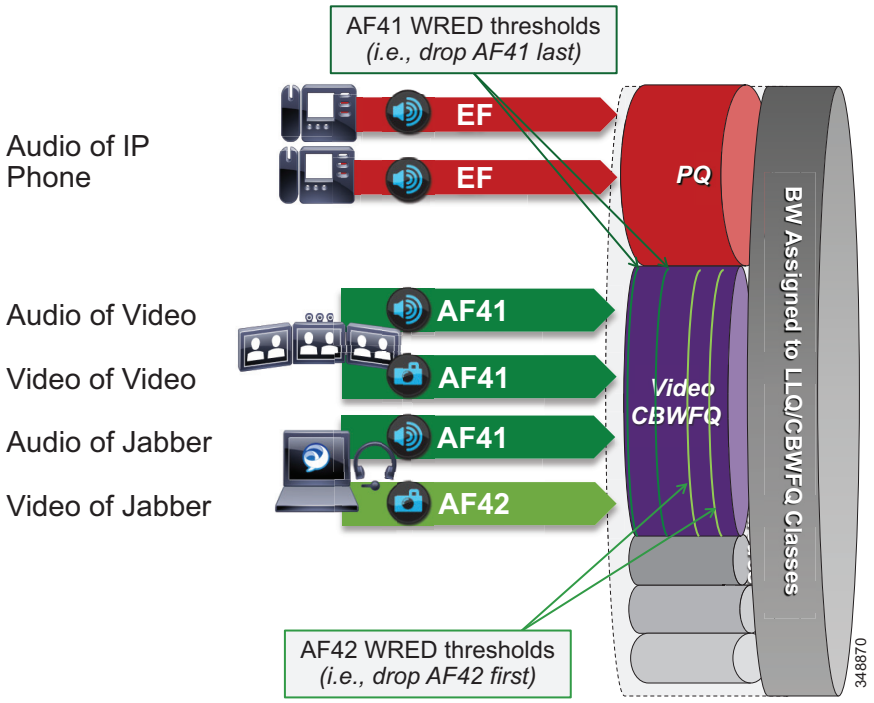
This section discusses the interface queuing. Figure 13-74 illustrates the voice PQ, the video CBWFQ, and the WRED thresholds used for the CBWFQ:

- Voice-only calls from trusted endpoints (EF) are mapped to the PQ.
- Prioritized video calls and Jabber share the same CBWFQ:
  - AF41 for audio and video streams of video calls from trusted endpoints
  - AF41 for audio streams of all calls from Jabber clients
  - AF42 for video streams of video calls from Jabber clients
- WRED is configured on the video queue:
  - Minimum and maximum thresholds for AF42:

    Approximately 10% to 30% of queue limit
  - Minimum and maximum thresholds for AF41:

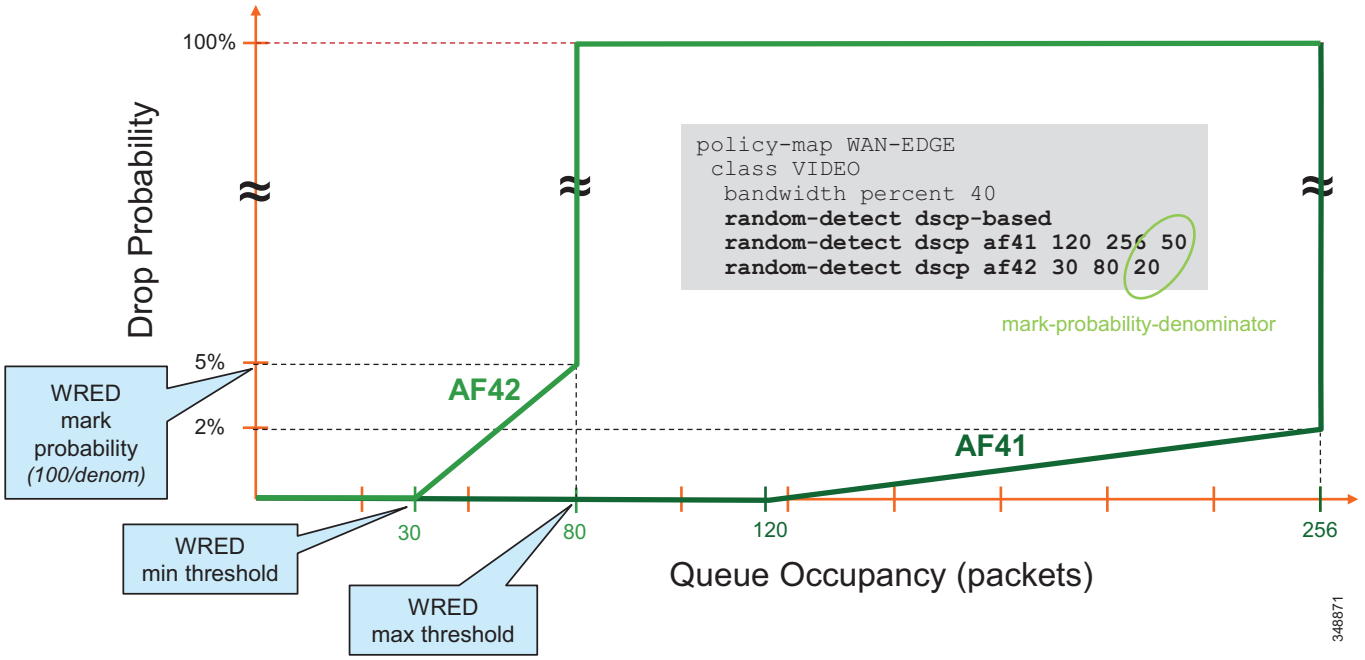
    Approximately 45% to 100% of queue limit

Figure 13-74 Queuing and Scheduling Collaboration Media



Weighted Random Early Detection (WRED) minimum and maximum thresholds are also configured in the Video CBWFQ. To illustrate how the WRED thresholds are configured, assume that the interface had been configured with a queue depth of 256 packets. Then following the guidelines listed above, the WRED minimum and maximum thresholds for AF42 and AF41 would be configured as illustrated in Figure 13-75.

Figure 13-75 Example of Video CBWFQ with WRED Thresholds



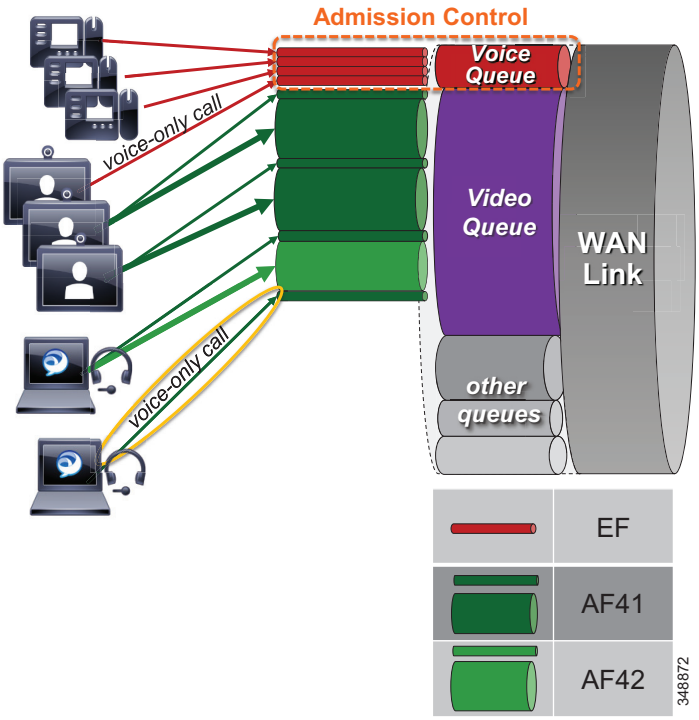
Provisioning and Admission Control

This section addresses admission control and provisioning bandwidth to the queues for each site type.

As mentioned previously, admission control is not used in this example case to manage the video bandwidth but instead to manage the audio traffic to ensure that the PQ is not over-subscribed. This is for voice-only calls.

Figure 13-76 illustrates the various call flows, their corresponding audio and video streams, and the queues to which they are directed.

Figure 13-76 Provisioning and Admission Control



The example in Figure 13-76 uses the following configuration:

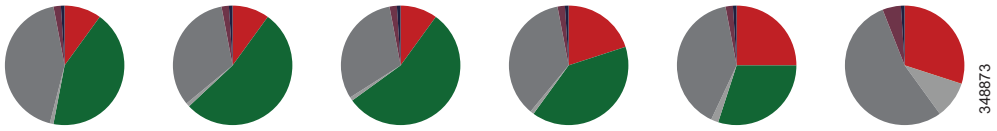
- Priority queue is provisioned for voice calls from trusted endpoints and is protected by admission control (ELCAC voice bandwidth pool).
- Video queue is over-provisioned for room-based video systems:
  - Ratios are applied to bandwidth usage for desktop video endpoints.
  - Jabber video calls can use any bandwidth unused by video room systems.
  - During congestion, video streams of Jabber calls are subject to WRED drops and dynamically reduce video bit rate.

Bandwidth Allocation Guidelines

The bandwidth allocations in Figure 13-77 are guidelines based solely on this Example Enterprise #1. They provide some guidance on percentages of available bandwidth for various common classes of Collaboration traffic. It is important to understand that bandwidth provisioning is highly dependent on utilization, and this will be different for each deployment and the user base being served at each site. The following examples provide a process to utilize for bandwidth provisioning. After provisioning the bandwidth, monitoring it and readjusting it are always necessary to ensure the best possible bandwidth provisioning and allocation necessary for an optimal user experience.

Figure 13-77 Bandwidth Allocation Guidelines

WAN Link Speed / Class	622 Mbps (OC12)	155 Mbps (OC3)	34-44 Mbps (E3/DS3)	10 Mbps	5 Mbps	<2 Mbps (T1/E1)
Control (%)	1	1	1	1	2	10
Voice (%)	10	10	10	20	25	30
Video (%)	43	53	55	40	30	--
Signalling (%)	2	2	2	2	2	5
Scavenger (%)	1	1	1	1	1	1
Default (%)	43	33	31	36	40	54



The following sections cover each site (Central, Large Branch, Small Branch, Micro Branch) and the link bandwidth provisioned for each class based on the number of users and available bandwidth for each class. Keep in mind that these values are based on bandwidth calculated for Layer 3 and above. Therefore, they do not include the Layer 2 overhead, which is dependent on the link type (Ethernet, Frame-relay, MPLS, and so forth). See the chapter on Network Infrastructure, page 3-1, for more information on Layer 2 overhead.

Central Site Link (100 Mbps) Bandwidth Calculation

As illustrated in Figure 13-78, the Central Site has the following bandwidth requirements:

- Voice queue (PQ): 10 Mbps (L3 bandwidth)

  125 calls @ G.711/G.722

- Unified CM Location link bandwidth for the voice pool:

  125 * 80 kbps = 10 Mbps

- Video queue: 55 Mbps (L3 bandwidth)
  - Immersive endpoint: 2 Mbps * 1 call = 2 Mbps
  - Video endpoints: 1.2 Mbps * 30 calls * 0.2 = 7.2 Mbps
  - TelePresence Servers: 1.5 Mbps * 40 calls * 0.5 = 30 Mbps
  - 55 Mbps – (2 Mbps + 7.2 Mbps + 30 Mbps) = 15.8 Mbps for Jabber media

    18 Jabber video calls @ 576p, or 50 @ 288p

    (Plus any leftover bandwidth)

Calculation Notes

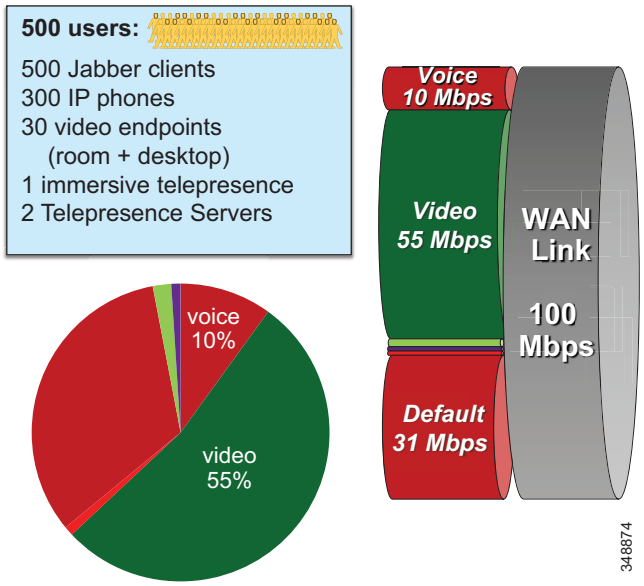
Immersive endpoints are sized for the busy hour. One endpoint is expected to be in a call across the WAN. This would be for a point-to-point call, since any conference call would terminate locally at the TelePresence server. It is important to take into account the worst-case scenario for the busy hour.

Video endpoints are sized for 20% WAN utilization (*0.2). A possible total of 30 calls at 1.2 Mbps is based on the number of endpoints. But assuming only 20% WAN utilization in active calls over the WAN, compared to active local calls, gives the WAN utilization rate of above 7.2 Mbps.

TelePresence Servers are sized at an average bit rate of 1.5 Mbps to account for the average of various endpoint resolutions from remote sites. The TelePresence Server would then be able to support up to 40 calls total (local and remote), and this is multiplied by 50% (0.5) to account for the possibility of half of the TelePresence calls going over the WAN while the other half might be serving local endpoints.

In addition there is 15.8 Mbps for Jabber calls, which could be 18 calls at 576p, or 50 calls at 288p, or variations thereof. This gives an idea of what the Jabber video calls have available for bandwidth. When more Jabber video calls occur past the 15.8 Mbps, packet loss will occur and will force all Jabber clients to adjust their bit rates down. This can be either a very subtle process with no visible user experience implications if the loss rate is low as new calls are added, or it can be very disruptive to the Jabber video if there is an immediate and sudden loss of packets. The expected packet loss rate as new video calls are added is helpful in determining the level of disruption in the user experience for this opportunistic class of video.

Figure 13-78 Central Site



Large Branch Link (34 Mbps) Bandwidth Calculation

As illustrated in Figure 13-79, the Large Branch site has the following bandwidth requirements:

- Voice queue (PQ): 3.4 Mbps (L3 bandwidth)

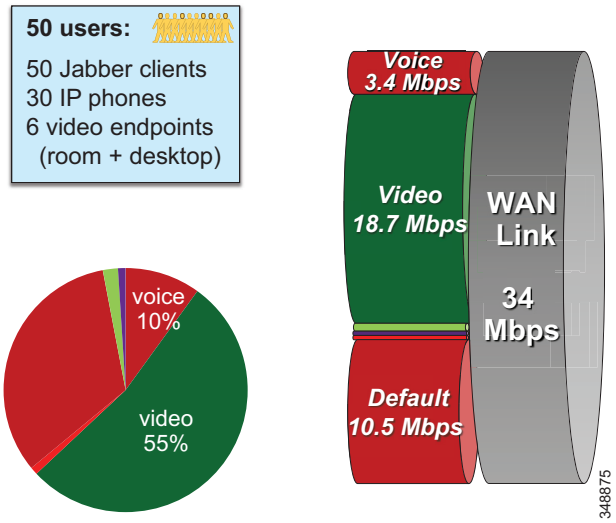
  42 calls @ G.711/G.722
- Unified CM Location link bandwidth for the voice pool:

  42 * 80 kbps = 3.360 Mbps
- Video queue: 18.7 Mbps (L3 bandwidth)
  - Video endpoints: 1.2 Mbps * 6 calls = 7.2 Mbps
  - 18.7 Mbps – 7.2 Mbps = 11.5 Mbps for Jabber media

    13 Jabber video calls @ 576p, or 36 @ 288p

    (Plus any leftover bandwidth)

Figure 13-79 Large Branch



Small Branch Link (10 Mbps) Bandwidth Calculation

As illustrated in Figure 13-80, the Small Branch site has the following bandwidth requirements:

- Voice queue (PQ): 2 Mbps (L3 bandwidth)

  25 calls @ G.711/G.722

- Unified CM Location link bandwidth for the voice pool:

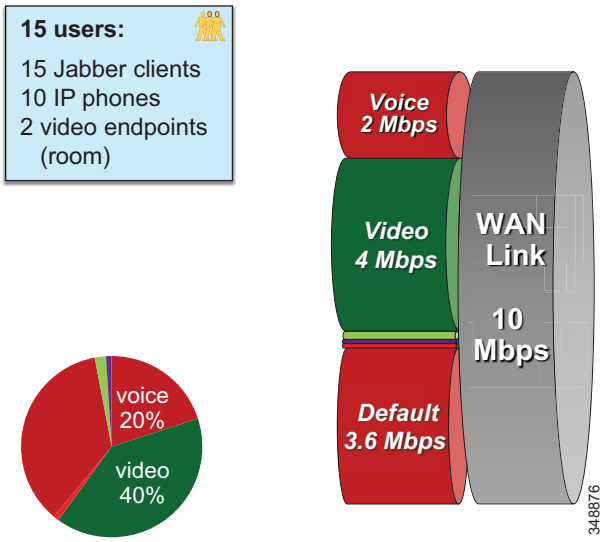
  25 * 80 kbps = 2 Mbps

- Video queue: 18.7 Mbps (L3 bandwidth)
  - Video endpoints: 1.2 Mbps * 2 calls = 2.4 Mbps
  - 4 Mbps – 2.4 Mbps = 1.6 Mbps for Jabber media

    2 Jabber video calls @ 576p, or 5 @ 288p

    (Plus any leftover bandwidth)

Figure 13-80 Small Branch

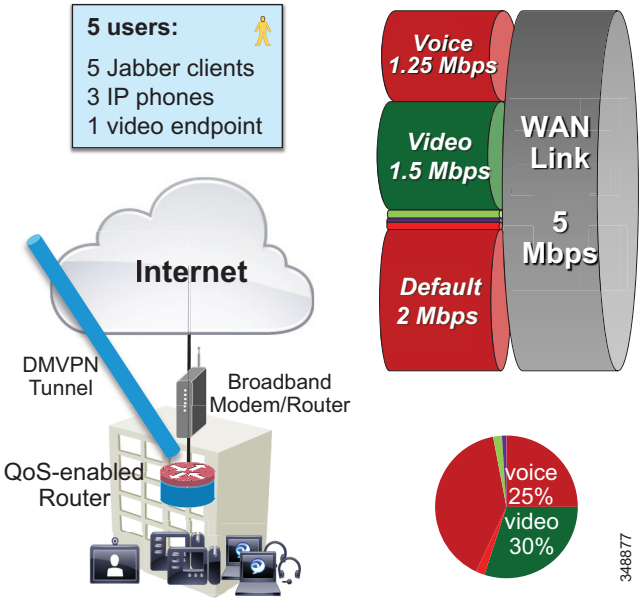


Micro Branch Broadband Internet Connectivity (5 Mbps) Bandwidth Calculation

As illustrated in Figure 13-81, the Micro Branch site has the following bandwidth requirements:

- Broadband Internet connectivity + DMVPN to central site
- Configure interface of VPN router to match broadband uplink speed
- Enable QoS on VPN router to prevent bufferbloat from TCP flows
- Asymmetric download/upload broadband: consider limiting transmit bit rate on video endpoint

Figure 13-81 Micro Branch

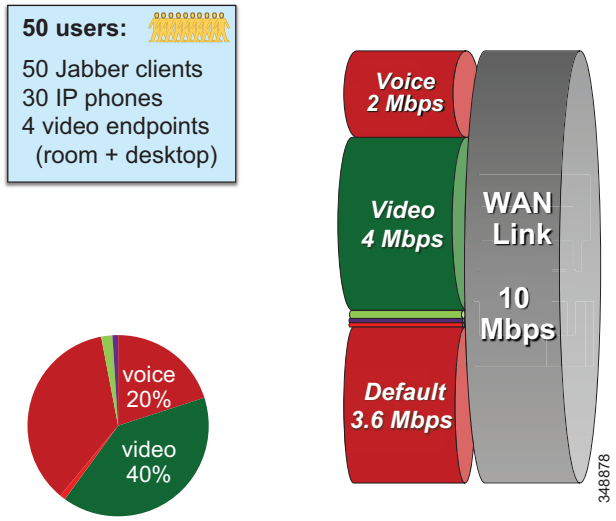


Large Branch with Constrained WAN Link (Enhanced Locations CAC Enabled for Video)

In specific branch sites with lower-speed WAN links, over-provisioning the video queue is not feasible (see Figure 13-82). ELCAC can be applied to these Location links for video to ensure that video calls do not over-subscribe the link bandwidth. This template requires using site-specific region configuration to limit maximum bandwidth used by video endpoints and Jabber clients. Also keep in mind that device mobility is required if Jabber users roam across sites.

Note Bandwidth for voice-only Jabber calls is subtracted from "voice" ELCAC, but it impacts the video queue (since it is marked AF41). Adjust the delta between video ELCAC bandwidth and video queue size.

Figure 13-82 Large Branch with Constrained WAN Link (Enhanced Locations CAC Enabled for Video)



As illustrated in Figure 13-82, a Large Branch site with a constrained WAN link (10 Mbps) has the following bandwidth requirements:

- Voice queue (PQ): 2 Mbps (L3 bandwidth)

  25 calls @ G.711/G.722

- Unified CM Location link bandwidth for the voice pool:

  25 * 80 kbps = 2 Mbps

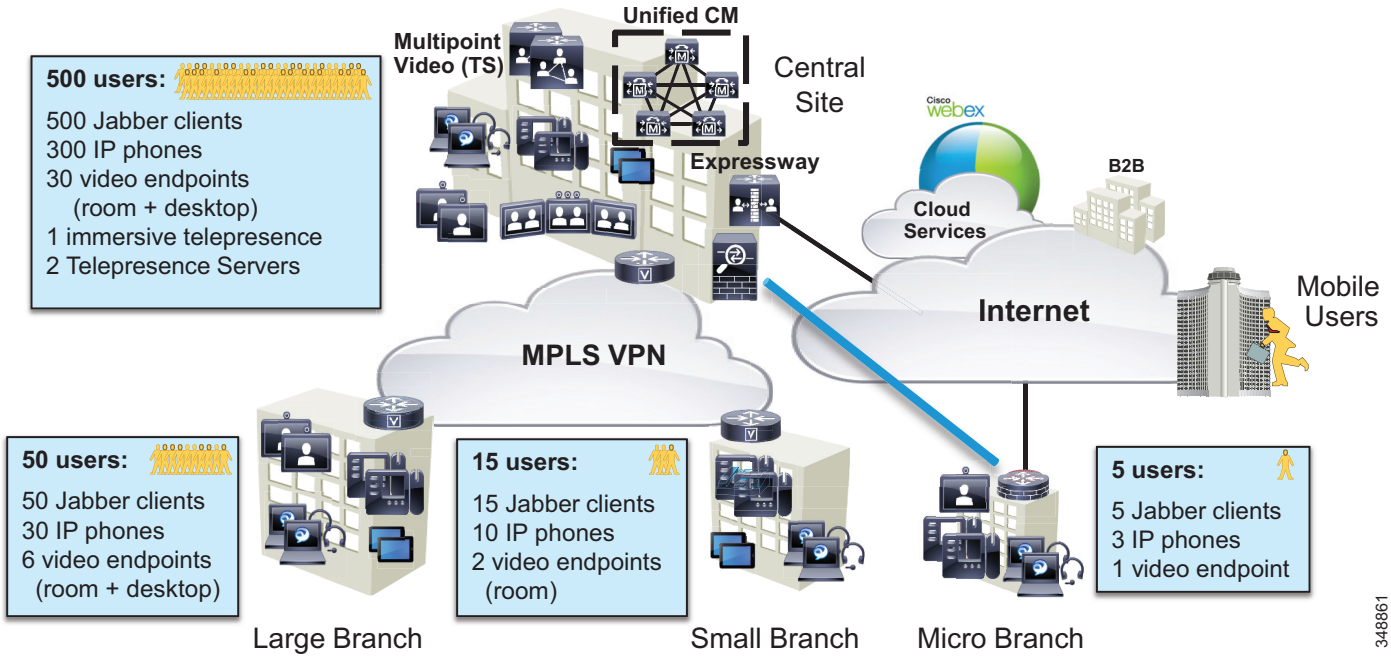
- Video queue: 4 Mbps (L3 bandwidth)
  - Possible usage: 2 calls @ 576p (768 kbps) + 5 calls @ 288p (320 kbps) = 3,136 kbps
  - Unified CM Location link bandwidth for video calls: 3.2 Mbps (L3 bandwidth)
  - Leaves room for L2 overhead, burstiness, and Jabber audio-only calls marked AF41

Example Enterprise #2

Example Enterprise #2 is a large enterprise with users across a large geographic area, with a data center (DC) at the headquarters site as well as multiple large, small, and micro-sized branches with roughly 500, 50, 15, and 5 users in each branch type, respectively. To simplify the illustration of the network, these categories of sites (HQ, large, small, and micro) are used as a template to size bandwidth considerations for each site that has a similar user base and endpoint density. Figure 13-83 illustrates each type of site. The enterprise has deployed Jabber with video to ensure that users have access to a video terminal for conferencing. The TelePresence video conferencing resources are located in the DC at HQ. IP phones are for voice-only communications; video endpoints are Jabber clients, Collaboration desktop endpoints (DX Series), and room endpoints (MX, Profile, and SX Series); and the HQ and large sites have immersive TelePresence units such as the IX Series.

Note Example Enterprise #2 is markedly different from Example Enterprise #1 in the sense that all endpoints (trusted and untrusted) in Example Enterprise #2 are configured to mark EF for all audio (voice-only and video calls) and mark video AF41 or AF42 for Jabber video. Also, Example Enterprise #2 uses Enhanced Locations CAC to protect the voice queue for the audio portion. Cisco Collaboration System Release (CSR) 12.*x* provides a feature whereby all audio can be deducted from the video pool. See the section on Enhanced Locations Call Admission Control, page 13-39, for more information.

Figure 13-83 Example Enterprise #2



The IT department is tasked with determining the bandwidth requirements for the WAN edge for each type of site in Example Enterprise #2. The following sections list the requirements and illustrate a methodology for applying QoS, determining bandwidth and queuing requirements, and determining admission control requirements.

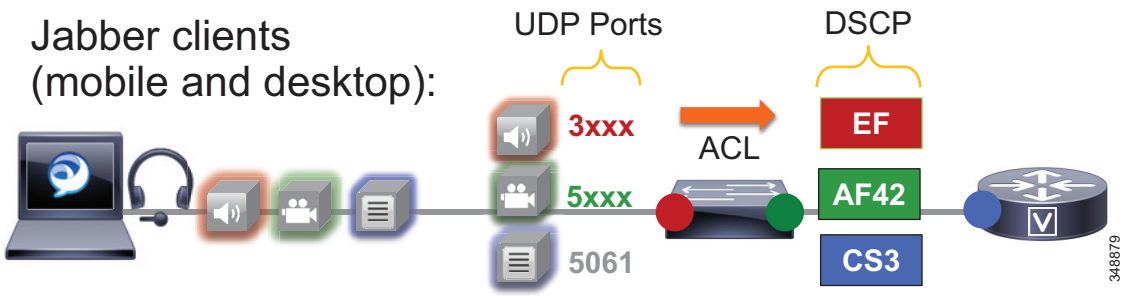
Identification and Classification

In this phase the QoS requirements are established across the enterprise.

Untrusted Endpoints (Jabber)

Jabber endpoints are untrusted and sit in the data VLAN. Example Enterprise #2 uses specific UDP port ranges to re-mark signaling and media at the access layer switch. In this case Unified CM is configured with a SIP Profile specifically for all Jabber clients to use the Separate Media and Signaling Port Range value of 3000 to 3999 for audio and 5000 to 5999 for video. The secure SIP signaling port of 5061 is used for Secure SIP signaling. This is illustrated in Figure 13-84.

Figure 13-84 Untrusted (Jabber) Endpoint QoS



The administrator creates an ACL for the access switches for the data VLAN to re-mark UDP ports to the following DSCP values:

- Audio: UDP Ports 3000 to 3999 marked to EF
- Video: UDP Ports 5000 to 5999 marked to AF42
- Signaling: TCP Ports 5060 to 5061 marked to CS3

Jabber classification summary:

- Audio streams of all Jabber calls (voice-only and video) are marked EF.
- Video streams of Jabber video calls are marked AF42.

For the Jabber endpoints we also recommend changing the default QoS values in the Jabber SIP profile. This is to ensure that, if for any reason the QoS is "trusted" via a wireless route or any other way, the correct "trusted" values will be the same as they would be for the re-marked value. Therefore, the QoS parameters in the SIP Profile need to be set as shown in Table 13-17.

Table 13-17 QoS Parameters in SIP Profile for Untrusted Jabber Endpoints

QoS Service Parameter Name (SIP Profile)	System Default Value	Changed Value
DSCP for Audio Calls	EF	
DSCP for Video Calls	AF41	AF42
DSCP for Audio Portion of Video Calls	AF41	EF
DSCP for TelePresence Calls	CS4	AF41
DSCP for Audio Portion of TelePresence Calls	CS4	EF

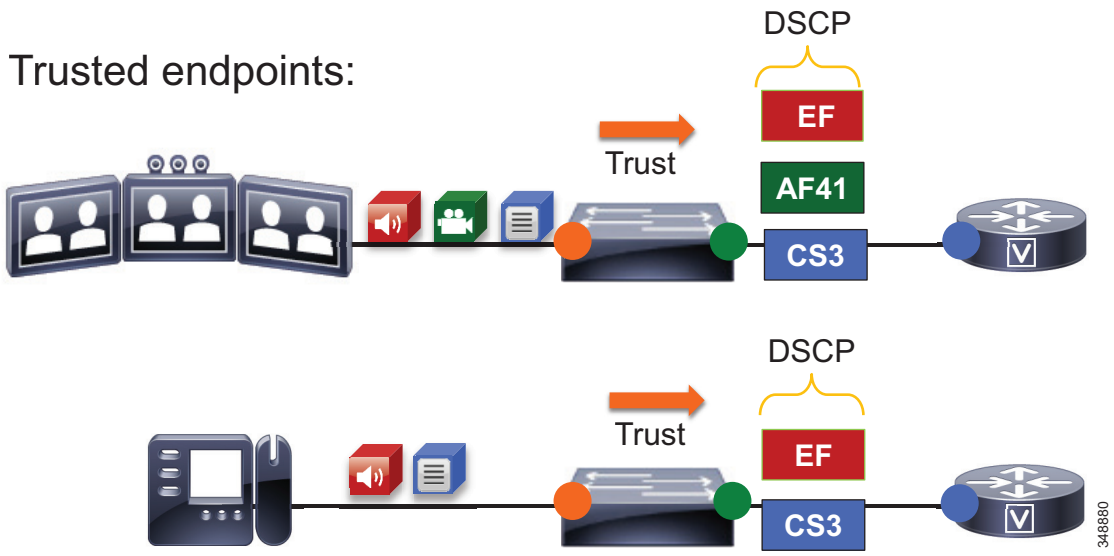
The configuration in Table 13-17 ensures that audio of Jabber clients will be set to EF and the video will be set to AF42, if for any reason they are trusted and not re-marked via UDP port range at the access switch. This is simply to ensure a consistent configuration across Jabber endpoints.

Trusted Endpoints

For the trusted endpoints, Cisco Discovery Protocol (CDP) is used and the QoS of the IP phones and video endpoint is trusted using the conditional trust mechanism configured at the access switch. The defaults need to be changed to ensure that all audio is set to EF for all endpoints. In this case Unified CM is configured with a SIP Profile that changes the audio of video and TelePresence calls to EF respectively.

Figure 13-85 illustrates the conditional trust (CDP based) and packet marking at the access switch.

Figure 13-85 Trusted Endpoint QoS



The administrator configures all access switches with a conditional QoS trust for IP phones and video and TelePresence endpoints, classified as follows:

- Audio streams of voice-only and video calls are marked EF.
- Video streams of video calls are marked AF41.

The administrator configures the Unified CM SIP Profile for trusted endpoints with the DSCP values listed in Table 13-18.

Table 13-18 QoS Parameters in SIP Profile for Trusted Endpoints

QoS Service Parameter Name (SIP Profile)	System Default Value	Changed Value
DSCP for Audio Calls	EF	
DSCP for Video Calls	AF41	
DSCP for Audio Portion of Video Calls	AF41	EF
DSCP for TelePresence Calls	CS4	AF41
DSCP for Audio Portion of TelePresence Calls	CS4	EF

At the WAN edge, on ingress it is expected that the packets arriving with a specific DSCP value have been trusted at the access layer or re-marked accordingly if they were not trusted at the access switch. As a failsafe practice, on ingress it is important to re-mark any untrusted traffic at the WAN edge that could not be re-marked at the access layer. While QoS is important in the LAN, it is paramount in the WAN; and as routers assume a trust on ingress traffic, it is important to configure the correct QoS policy that aligns with the business requirements and user experience. The WAN edge re-marking is always done on the ingress interface into the router, while the queuing and scheduling is done on the egress interface. The following example walks through the WAN ingress QoS policy as well as the egress queuing policy. Figure 13-86 illustrates the configuration and the re-marking process.

In Figure 13-86 the packets from both the trusted and untrusted areas of the network are identified and classified with the appropriate DSCP marking via the trust methods discussed or via a simple ACL matching on UDP port ranges. Keep in mind that this ACL could also match more granularly on IP address or some other attributes that would further limit the scope of the marking.

Figure 13-86 Example Router Ingress QoS Policy Process – Step 1

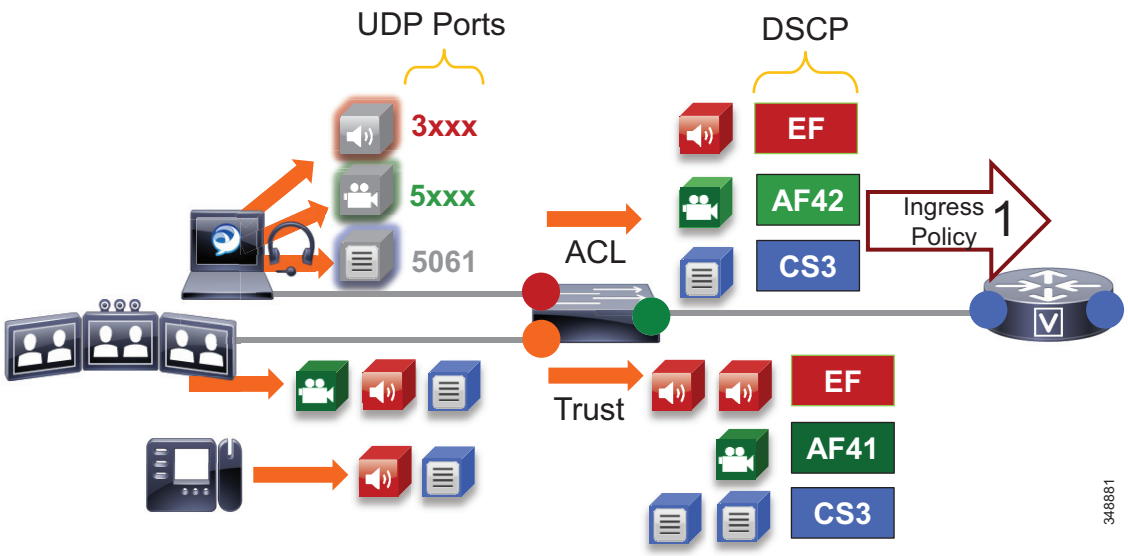
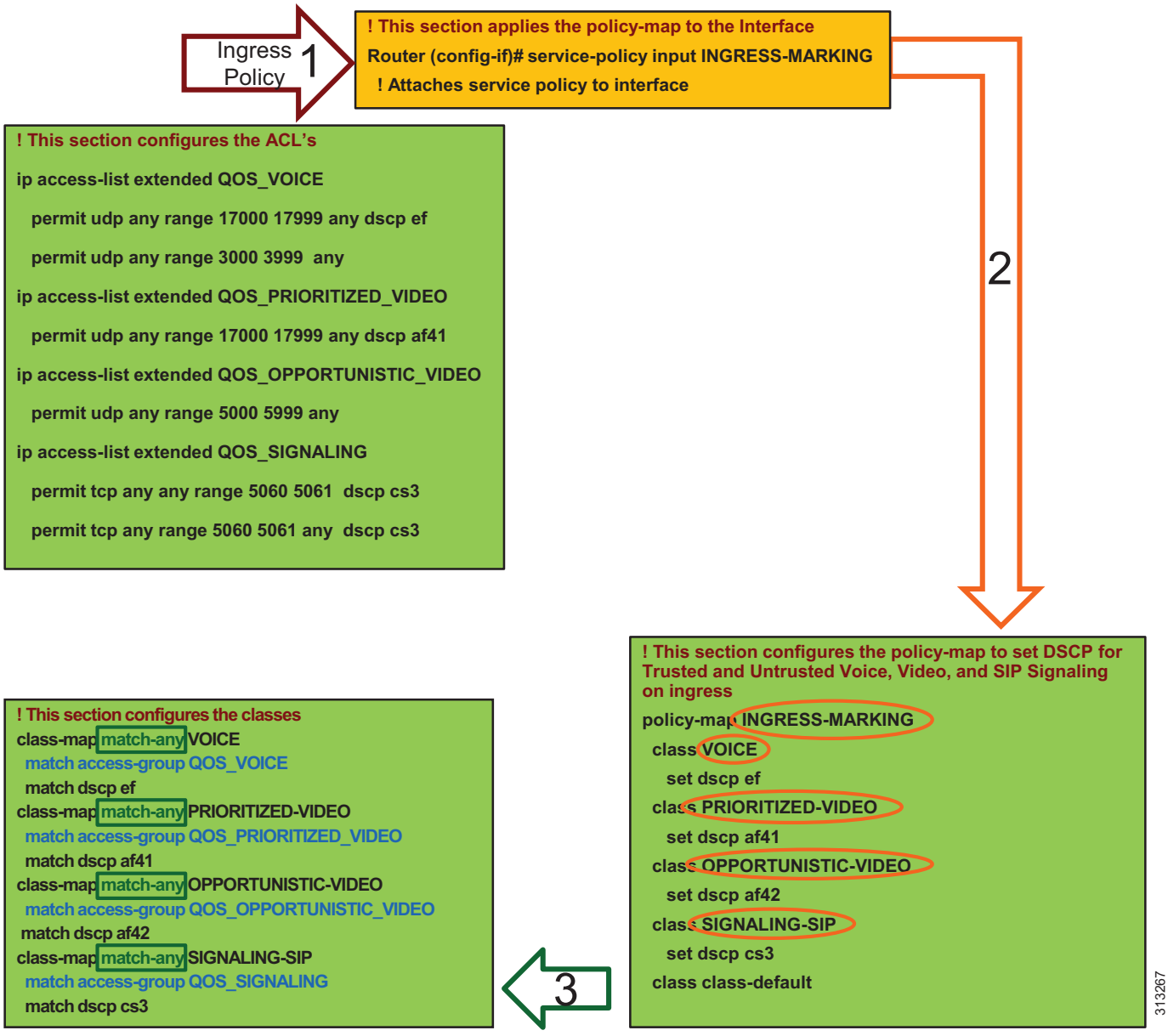


Figure 13-87 through illustrate the policy matching criteria and DSCP re-marking. The process involves the following steps shown in the figures:

1. Packets arrive into the router ingress interface, which is configured with an input service policy (Figure 13-87).

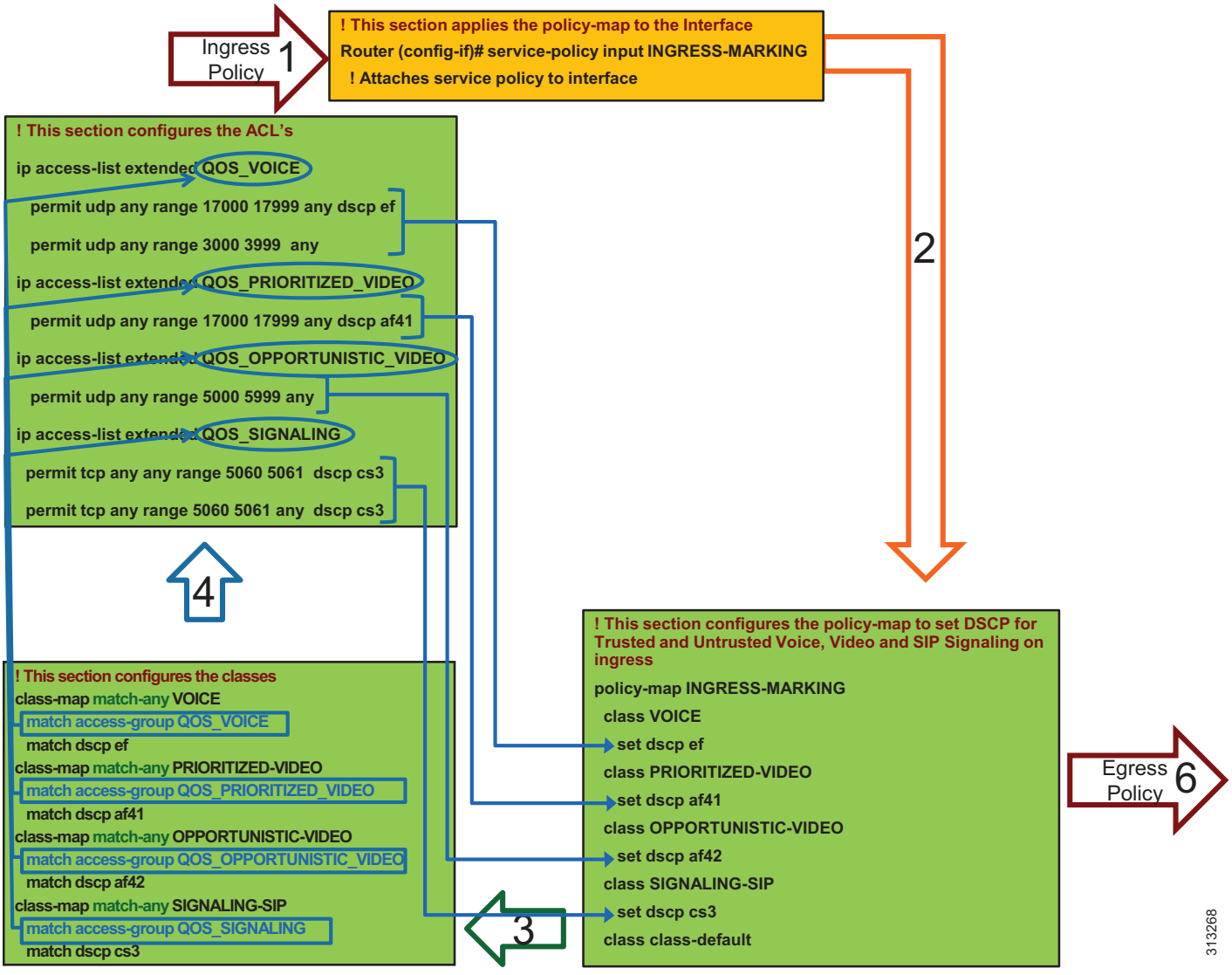
2. The policy-map is configured with 4 classes of traffic setting the appropriate DSCP: VOICE setting a DSCP of EF, PRIORITIZED-VIDEO setting a DSCP of AF41, OPPORTUNISTIC-VIDEO setting a DSCP of AF42, and SIGNALING-SIP setting a DSCP of CS3 (Figure 13-87).
3. Each one of these classes matches a class-map of the same name configured with **match-any** criteria and a DSCP match as well as an ACL match. This match-any criteria means that the process will start top-down, and the first matching criteria will be executed and thus set the DSCP according to each class in the policy-map statements. Another option is **match-all**, which would require all criteria to be matched and thus would match DSCP *and* ACL. This, however, would not provide the intended functionality of re-marking either marked *or* unmarked traffic.

Figure 13-87 Example Router Ingress QoS Policy Process – Steps 1 to 3



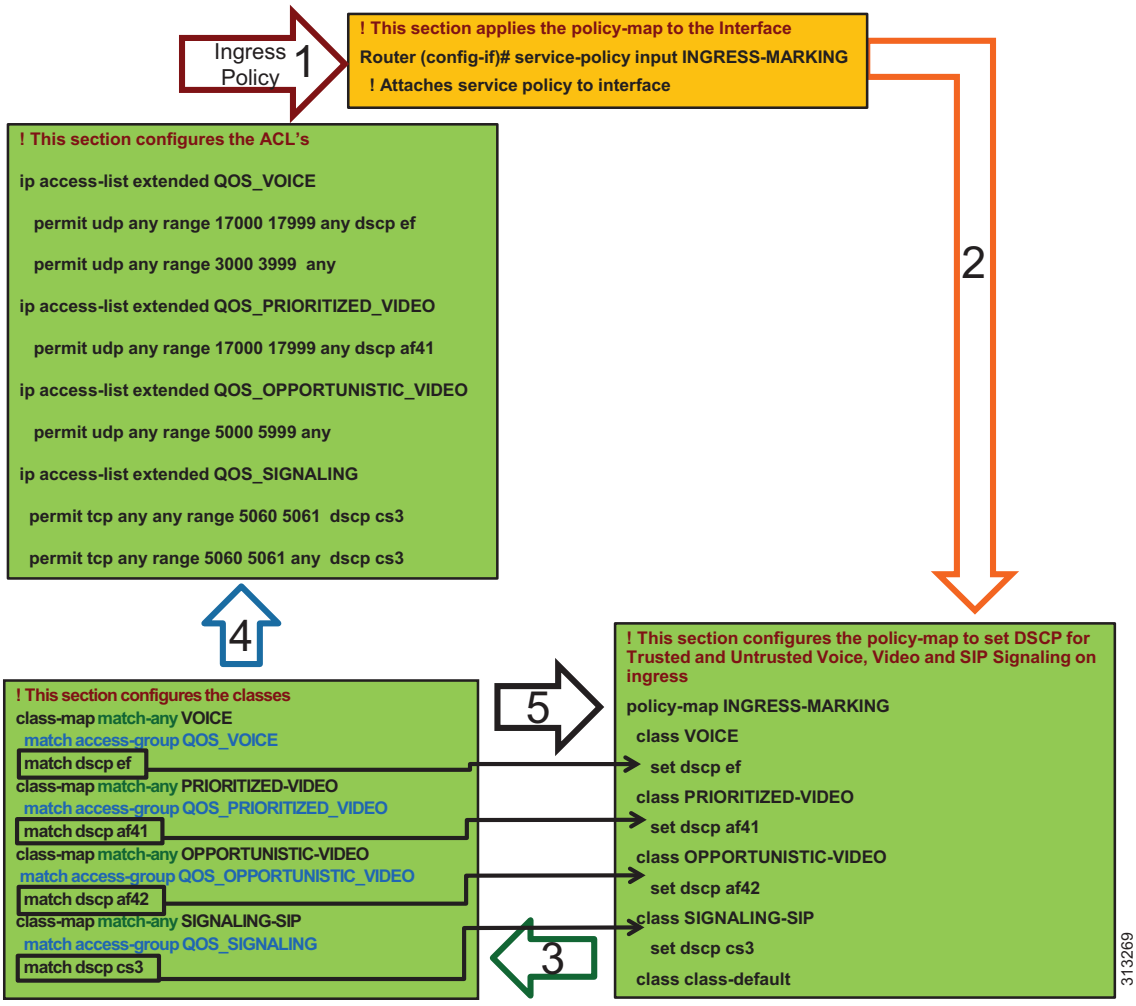
4. In step 4, the first line in the class-map statement is parsed, which is the ACL that matches the UDP ports set in Unified CM in the Identification and Classification section. When the ACL criteria are met (protocol, port range, and in some cases DSCP), then the traffic is marked as is configured in the corresponding policy-map statements (Figure 13-88). Note that Jabber Audio is marked EF and Jabber Video is marked AF42, in line with the policy in Figure 13-86.

Figure 13-88 Example Router Ingress QoS Policy Process – Step 4



5. In step 5, the traffic that did not match the first statement goes to the next match statement in the class-map, which is **match dscp** (Figure 13-89). If the traffic simply matches the DSCP, then DSCP is set again to the same value that was matched and as is configured in the policy-map statements. In this case the router is simply matching on DSCP and resetting the DSCP to the same value. This is a catch-all setting for the trusted DSCP from servers and applications coming into the WAN router.

Figure 13-89 Example Router Ingress QoS Policy Process – Step 5

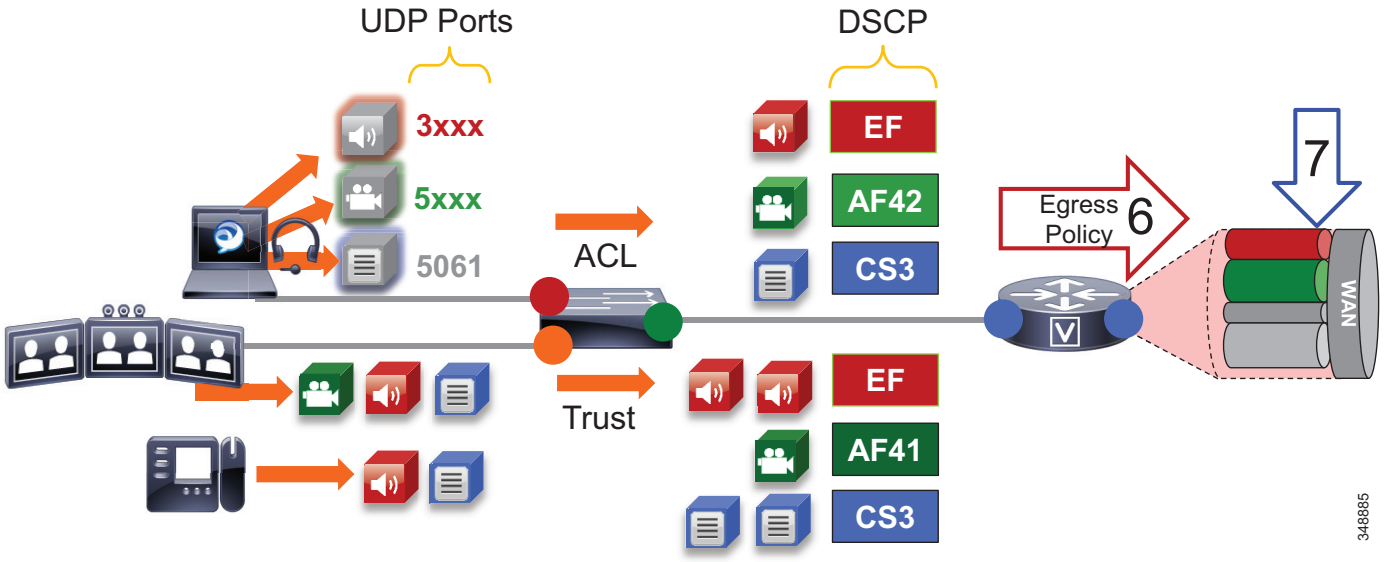


Note This is an example QoS ingress marking policy based on the Cisco Common Classification Policy Language (C3PL). Refer to your specific router configuration guide for information on how to achieve a similar policy on a Cisco router supporting C3PL and for any updated commands.

6. The traffic goes to an outbound interface to be queued and scheduled by an output service policy that has 3 queues created: a Priority Queue called VOICE, a CBWFQ called VIDEO, and another CBWFQ called SIGNALING (Figure 13-90. This highlights the fact that this egress queuing policy is based only on DSCP as network marking occurring at the access switch and/or on ingress into the

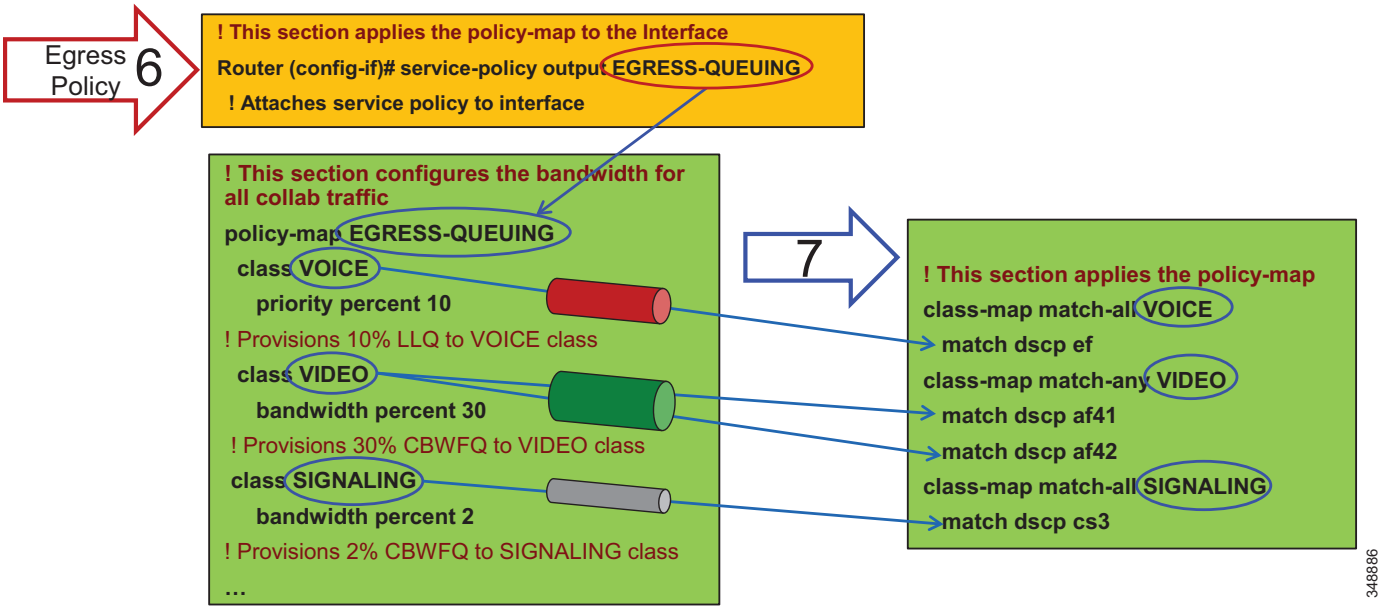
WAN router ingress interface. This is an example simply to illustrate the matching criteria and queues, and it does not yet contain the WRED functionality (covered in the next subsection). For more information on WRED, see the next section on WAN Queuing and Scheduling, page 13-117.

Figure 13-90 Example Router Egress Queuing Policy Process – Step 6



7. The traffic is matched against the class-map match statements, and all traffic marked EF goes to the VOICE PQ, AF41 and AF42 traffic goes to the VIDEO CBWFQ, and CS3 traffic goes to the SIGNALING CBWFQ (Figure 13-91).

Figure 13-91 Example Router Egress Queuing Policy Process – Step 7



Note This is an example egress queuing policy based on the Cisco Common Classification Policy Language (C3PL). Refer to your specific router configuration guide for information on how to achieve a similar policy on a Cisco router supporting C3PL and for any updated commands.

WAN Queuing and Scheduling

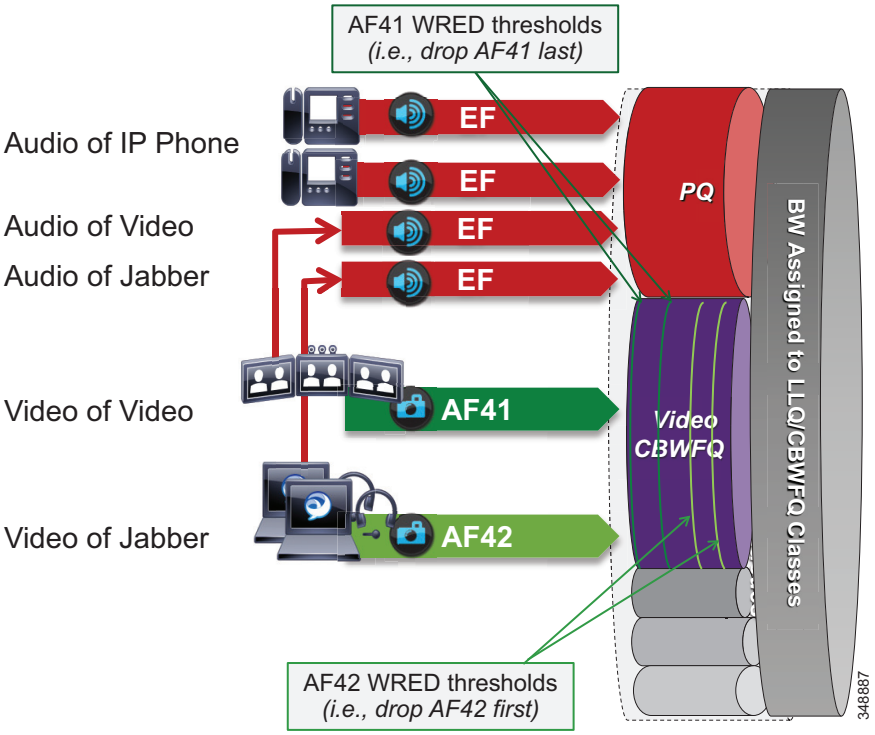
This section discusses the interface queuing. Figure 13-92 illustrates the voice PQ, video CBWFQ, and WRED thresholds used for the CBWFQ:

- All audio from all endpoints (trusted and untrusted) marked EF is mapped to the PQ.
- Video calls and Jabber share the same CBWFQ:
  - EF for audio streams of video calls from trusted endpoints
  - AF41 for video streams of video calls from trusted endpoints
  - EF for audio streams of all calls from Jabber clients
  - AF42 for video streams of video calls from Jabber clients
- WRED is configured on the video queue:
  - Minimum and maximum thresholds for AF42:

    Approximately 10% to 30% of queue limit
  - Minimum and maximum thresholds for AF41:

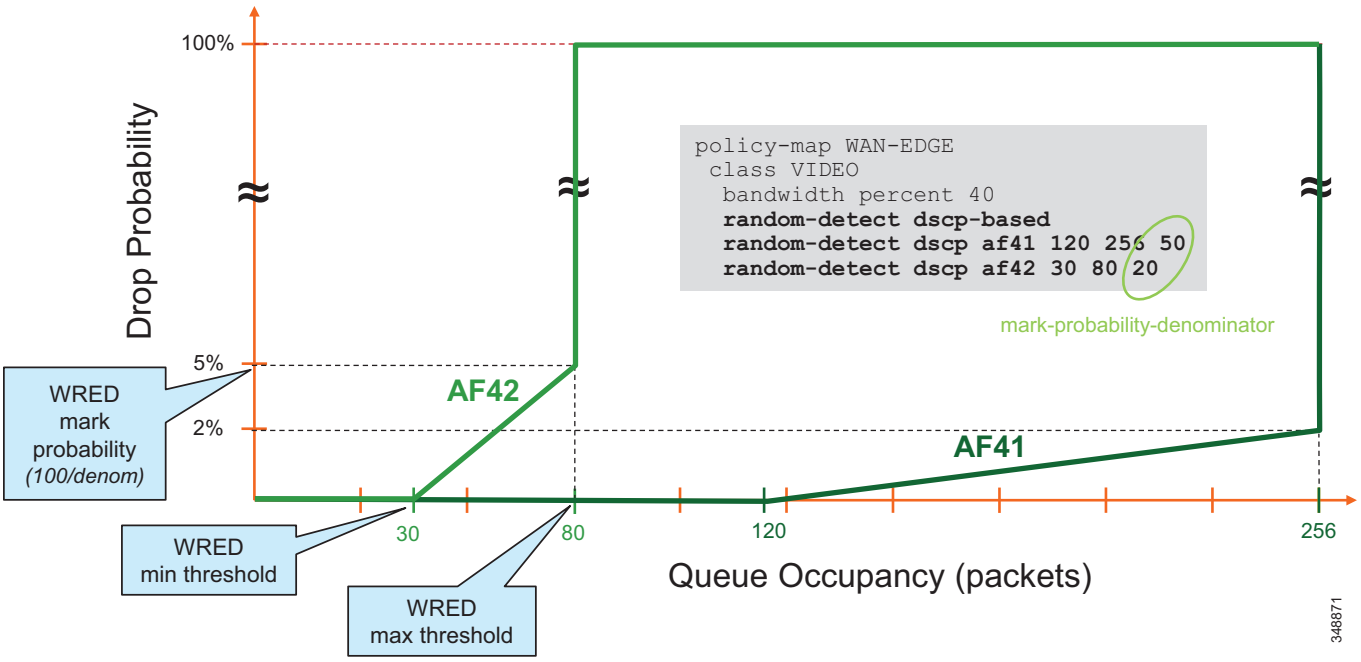
    Approximately 45% to 100% of queue limit

Figure 13-92 Queuing and Scheduling Collaboration Media



Weighted Random Early Detection (WRED) minimum and maximum thresholds are also configured in the Video CBWFQ. To illustrate how the WRED thresholds are configured, assume that the interface had been configured with a queue depth of 256 packets. Then following the guidelines listed above, the WRED minimum and maximum thresholds for AF42 and AF41 would be configured as illustrated in Figure 13-93.

Figure 13-93 Example of Video CBWFQ with WRED Threshold



Provisioning and Admission Control

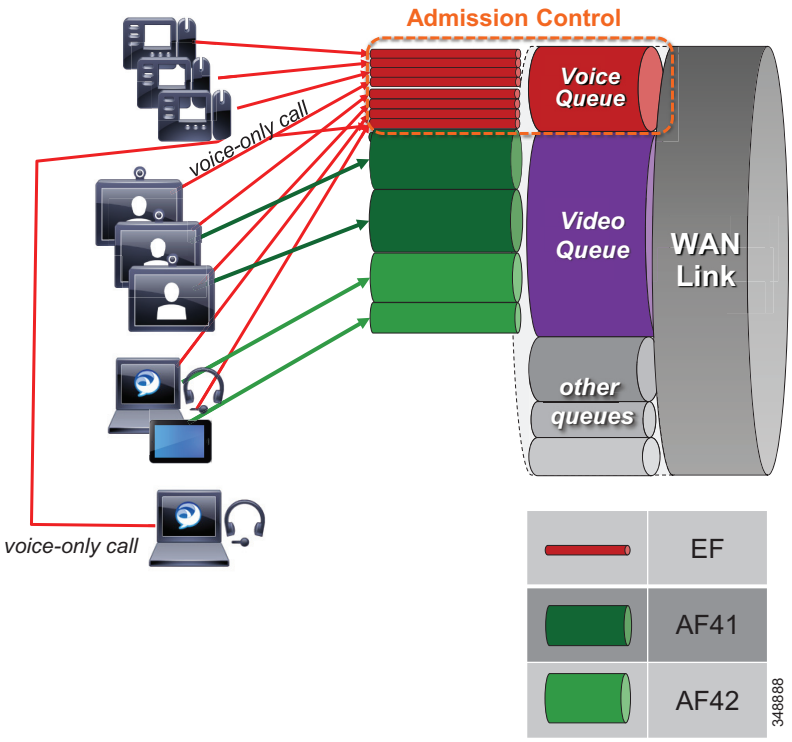
This section addresses admission control and provisioning bandwidth to the queues for each site type.

As mentioned previously, admission control is not used in this example case to manage the video bandwidth but instead to manage the audio traffic to ensure that the PQ is not over-subscribed. And in this Example Enterprise #2, the voice pool in Enhanced Locations CAC will be admitting the audio for both the voice-only calls and the video calls.

In Unified CM this feature is enabled by setting the service parameter **Deduct Audio Bandwidth from Audio Pool for Video Call** to **True** under the Call Admission Control section of the CallManager service called. False is the default setting, and by default Unified CM deducts both audio and video streams of video calls from the video pool. This parameter will change that behavior and is key to the QoS alterations in Example Enterprise #2.

Figure 13-94 illustrates the various call flows, their corresponding audio and video streams, and the queues to which they are directed.

Figure 13-94 Provisioning and Admission Control



The example in Figure 13-94 uses the following configuration:

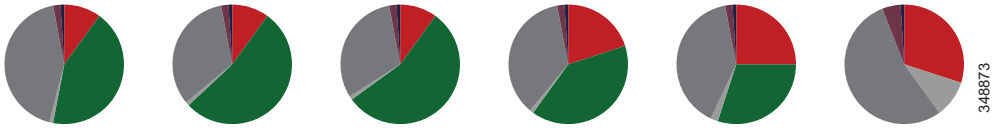
- Priority queue is provisioned for all calls from both trusted and untrusted endpoints, and it is protected by admission control (E-LCAC voice BW pool).
- Video queue is over-provisioned for room-based video systems:
  - Ratios are applied to bandwidth usage for desktop video endpoints.
  - Jabber video calls can use any bandwidth unused by video room systems.
  - During congestion, video streams of Jabber calls are subject to WRED drops and dynamically reduce video bit rate.

Bandwidth Allocation Guidelines

The bandwidth allocations in Figure 13-95 are guidelines based solely on this Example Enterprise #2. They provide some guidance on percentages of available bandwidth for various common classes of Collaboration traffic. It is important to understand that bandwidth provisioning is highly dependent on utilization, and this will be different for each deployment and the user base being served at each site. The following examples provide a process to utilize for bandwidth provisioning. After provisioning the bandwidth, monitoring it and readjusting it are always necessary to ensure the best possible bandwidth provisioning and allocation necessary for an optimal user experience.

Figure 13-95 Bandwidth Allocation Guidelines

WAN Link Speed / Class	622 Mbps (OC12)	155 Mbps (OC3)	34-44 Mbps (E3/DS3)	10 Mbps	5 Mbps	<2 Mbps (T1/E1)
Control (%)	1	1	1	1	2	10
Voice (%)	10	10	10	20	25	30
Video (%)	43	53	55	40	30	--
Signalling (%)	2	2	2	2	2	5
Scavenger (%)	1	1	1	1	1	1
Default (%)	43	33	31	36	40	54



The following sections describe each site (Central, Large Branch, Small Branch, and Micro Branch) and the link bandwidth provisioned for each class based on the number of users and available bandwidth for each class. Keep in mind that these values are based on bandwidth calculated for Layer 3 and above. Therefore, they do not include the Layer 2 overhead, which is dependent on the link type (Ethernet, Frame-relay, MPLS, and so forth). See the chapter on Network Infrastructure, page 3-1, for more information on Layer 2 overhead.

Also, note that the audio portion of bandwidth for video calls is now deducted from the voice pool. So when provisioning the voice queue, this will include the audio bandwidth for both voice-only and video calls. These examples are the same as those for Example Enterprise #1, page 13-91. The only difference is that for Example Enterprise #2 the audio portion of bandwidth for video calls is deducted from the voice admission control pool, and the audio streams go into the voice queue.

Central Site Link (100 Mbps) Bandwidth Calculation

As illustrated in Figure 13-96, the Central Site has the following bandwidth requirements:

- Voice queue (PQ): 10 Mbps (L3 bandwidth)

  125 calls @ G.711/G.722
- Unified CM Location link bandwidth for the voice pool:

  125 * 80 kbps = 10 Mbps
- Video queue: 55 Mbps (L3 bandwidth)
  - Immersive endpoint: 2 Mbps * 1 call = 2 Mbps
  - Video endpoints: 1.2 Mbps * 30 calls * 0.2 = 7.2 Mbps
  - TelePresence Servers: 1.5 Mbps * 40 calls * 0.5 = 30 Mbps
  - 55 Mbps – (2 Mbps + 7.2 Mbps + 30 Mbps) = 15.8 Mbps for Jabber media

    18 Jabber video calls @ 576p, or 50 @ 288p

    (Plus any leftover bandwidth)

Calculation Notes

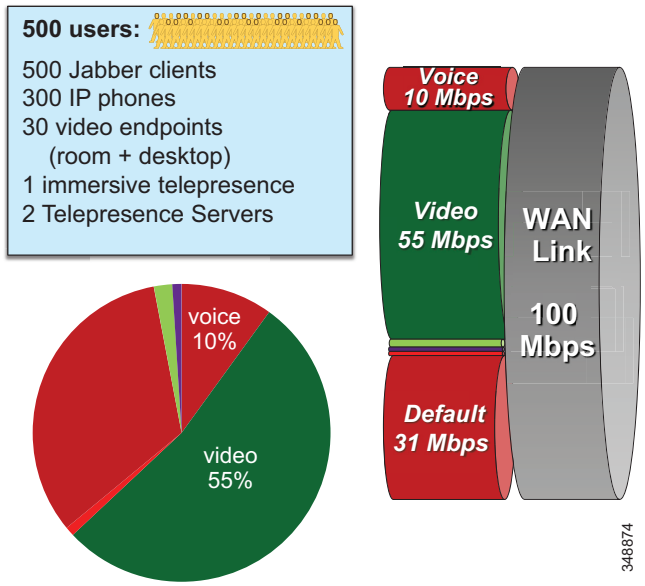
Immersive endpoints are sized for the busy hour. One endpoint is expected to be in a call across the WAN. This would be for a point-to-point call, since any conference call would terminate locally at the TelePresence server. It is important to take into account the worst-case scenario for the busy hour.

Video endpoints are sized for 20% WAN utilization (*0.2). A possible total of 30 calls at 1.2 Mbps is based on the number of endpoints. But assuming only 20% WAN utilization in active calls over the WAN, compared to active local calls, gives the WAN utilization rate of above 7.2 Mbps.

TelePresence Servers are sized at an average bit rate of 1.5 Mbps to account for the average of various endpoint resolutions from remote sites. The TelePresence Server would then be able to support up to 40 calls total (local and remote), and this is multiplied by 50% (0.5) to account for the possibility of half of the TelePresence calls going over the WAN while the other half might be serving local endpoints.

In addition there is 15.8 Mbps for Jabber calls, which could be 18 calls at 576p, or 50 calls at 288p, or variations thereof. This gives an idea of what the Jabber video calls have available for bandwidth. When more Jabber video calls occur past the 15.8 Mbps, packet loss will occur and will force all Jabber clients to adjust their bit rates down. This can be either a very subtle process with no visible user experience implications if the loss rate is low as new calls are added, or it can be very disruptive to the Jabber video if there is an immediate and sudden loss of packets. The expected packet loss rate as new video calls are added is helpful in determining the level of disruption in the user experience for this opportunistic class of video.

Figure 13-96 Central Site



Large Branch Link (34 Mbps) Bandwidth Calculation

As illustrated in Figure 13-97, the Large Branch site has the following bandwidth requirements:

- Voice queue (PQ): 3.4 Mbps (L3 bandwidth)

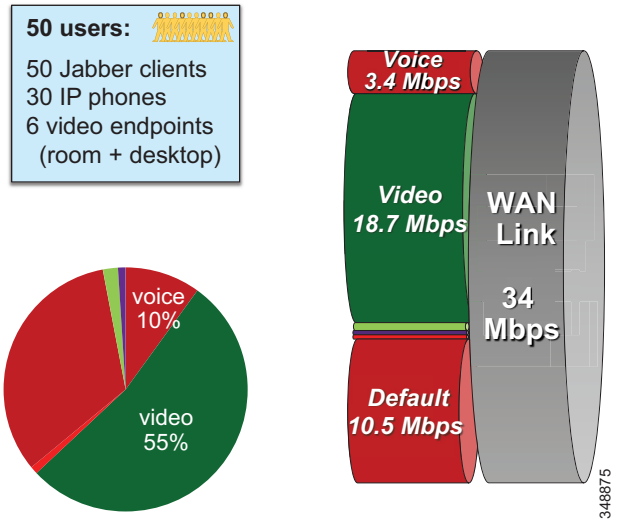
  42 calls @ G.711/G.722
- Unified CM Location link bandwidth for the voice pool:

  42 * 80 kbps = 3.360 Mbps
- Video queue: 18.7 Mbps (L3 bandwidth)
  - Video endpoints: 1.2 Mbps * 6 calls = 7.2 Mbps
  - 18.7 Mbps – 7.2 Mbps = 11.5 Mbps for Jabber media

    13 Jabber video calls @ 576p, or 36 @ 288p

    (Plus any leftover bandwidth)

Figure 13-97 Large Branch



Small Branch Link (10 Mbps) Bandwidth Calculation

As illustrated in Figure 13-98, the Small Branch site has the following bandwidth requirements:

- Voice queue (PQ): 2 Mbps (L3 bandwidth)

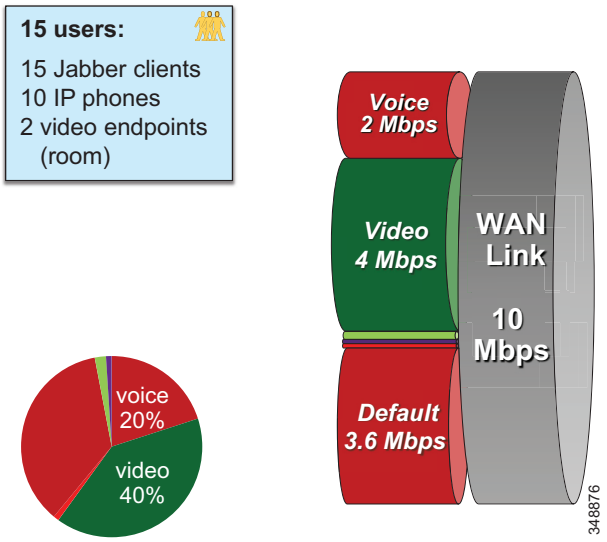
  25 calls @ G.711/G.722
- Unified CM Location link bandwidth for the voice pool:

  25 * 80 kbps = 2 Mbps
- Video queue: 18.7 Mbps (L3 bandwidth)
  - Video endpoints: 1.2 Mbps * 2 calls = 2.4 Mbps
  - 4 Mbps – 2.4 Mbps = 1.6 Mbps for Jabber media

    2 Jabber video calls @ 576p, or 5 @ 288p

    (Plus any leftover bandwidth)

Figure 13-98 Small Branch

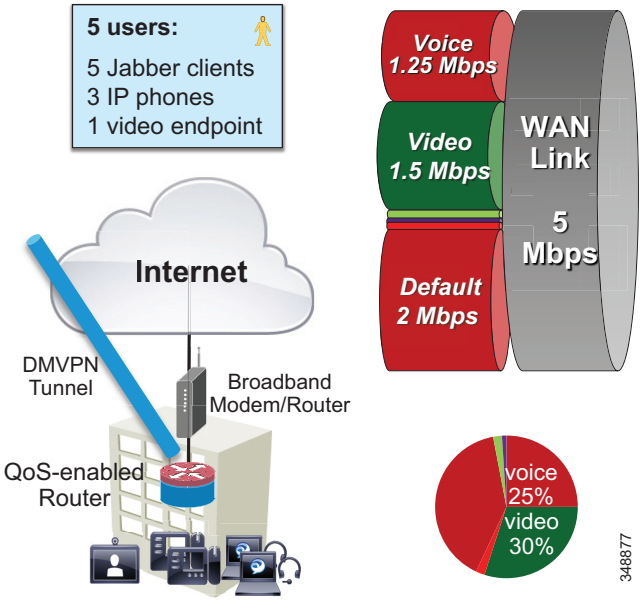


Micro Branch Broadband Internet Connectivity (5 Mbps) Bandwidth Calculation

As illustrated in Figure 13-99, the Micro Branch site has the following bandwidth requirements:

- Broadband Internet connectivity + DMVPN to central site
- Configure interface of VPN router to match broadband uplink speed
- Enable QoS on VPN router to prevent bufferbloat from TCP flows
- Asymmetric download/upload broadband: consider limiting transmit bit rate on video endpoint

Figure 13-99 Micro Branch

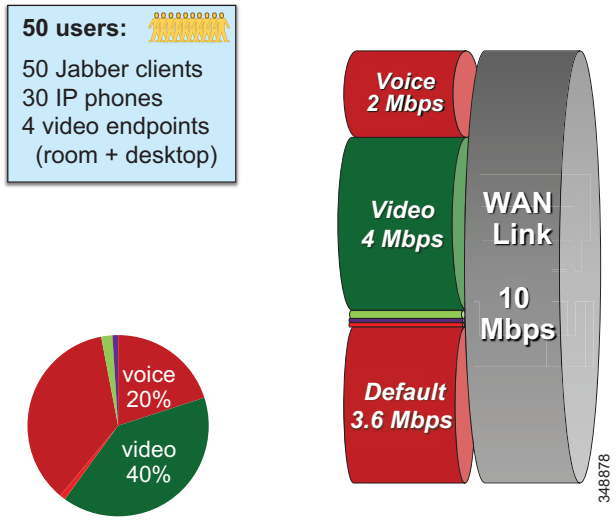


Large Branch with Constrained WAN Link (Enhanced Locations CAC Enabled for Video)

In specific branch sites with lower-speed WAN links, over-provisioning the video queue is not feasible (see Figure 13-100). ELCAC can be applied to these Location links for video to ensure that video calls do not over-subscribe the link bandwidth. This template requires using site-specific region configuration to limit maximum bandwidth used by video endpoints and Jabber clients. Also keep in mind that device mobility is required if Jabber users roam across sites.

Note Because audio bandwidth for both voice-only and video calls is deducted from the voice CAC pool, there is no need for any queue bandwidth adjustment as is the case in Example Enterprise #1.

Figure 13-100 Large Branch with Constrained WAN Link (Enhanced Locations CAC Enabled for Video)



As illustrated in Figure 13-100, a Large Branch site with a constrained WAN link (10 Mbps) has the following bandwidth requirements:

- Voice queue (PQ): 2 Mbps (L3 bandwidth)

  25 calls @ G.711/G.722

- Unified CM Location link bandwidth for the voice pool:

  25 * 80 kbps = 2 Mbps

- Video queue: 4 Mbps (L3 bandwidth)
  - Possible usage: 2 calls @ 576p (768 kbps) + 5 calls @ 288p (320 kbps) = 3,136 kbps
  - Unified CM Location link bandwidth for video calls: 3.2 Mbps (L3 bandwidth)
  - Leaves room for L2 overhead, burstiness, and Jabber audio-only calls marked AF41

Dial Plan

Revised: March 1, 2018

The dial plan is one of the key elements of a Unified Communications and Collaboration system, and an integral part of all call processing agents. Generally, the dial plan is responsible for instructing the call processing agent on how to route calls. Specifically, the dial plan performs the following main functions:

- Endpoint addressing

  For destinations registered with the call processing agent, addresses are assigned to provide reachability. These internal destinations include all endpoints (such as IP phones, video endpoints, soft clients and analog endpoints) and applications (such as voicemail systems, auto attendants, and conferencing systems).

- Path selection

  Depending on the calling device and the destination dialed, a path to the dialed destination is selected. If a secondary path is available, this path will also be considered if the primary path fails.

- Calling privileges

  Different groups of devices can be assigned to different classes of service, by granting or denying access to certain destinations. For example, lobby phones might be allowed to reach only internal and local PSTN destinations, while executive phones could have unrestricted PSTN access.

- Manipulation of dialed destination

  On the path from the dialing device to the dialed destination, the dial plan can apply manipulations to the dialed destination. For example, users in the US might dial 9011496901234 to reach a destination in the PSTN in Germany, while a user in France might be able to reach the same destination by dialing 000496901234. This dialed destination would need to be presented as 011496901234 to a PSTN trunk on a gateway in the US and as 00496901234 to a PSTN trunk on a gateway in France.

- Presentation of information about identities involved in the call

  During session establishment and also while in the call, on both the calling and the called device, information about the other device is displayed. Depending on call state and direction, this includes calling, diverting, alerting, and connected party information. The dial plan can define mappings that influence the format and content of information displayed.

This chapter presents information intended to guide the system designer toward a dial plan that accommodates the legacy dialing habits of telephony and video users, while also taking advantage of new functionality afforded by the increasing integration between computing technology and telephony, such as dialing from contacts, click-to-call actions from computers and smart phones, and adoption of mobility-related features. The chapter is structured to offer information about the following main areas:

- Dial Plan Fundamentals, page 14-3

  This section covers general concepts commonly used in enterprise voice and video dial plans.

- Dial Plan Elements, page 14-13

  This section introduces the various dial plan elements available in the architectural elements of an enterprise collaboration solution, including Cisco Unified Communications Manager (Unified CM) and Cisco TelePresence Video Communication Server (VCS).

- Recommended Design, page 14-56

  This section contains design guidelines related to multisite collaboration networks, endpoint addressing, and building classes of service. Also, dial plan integration between Unified CM and VCS is covered.

- Special Considerations, page 14-79

For more details, refer to the *System Configuration Guide for Cisco Unified Communications Manager*, the *Feature Configuration Guide for Cisco Unified Communications Manager*, the Cisco IOS Voice and Video Configuration guides, and other product documentation available at

https://www.cisco.com

What's New in This Chapter

Table 14-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 14-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Do not assign a single route filter to too many route patterns.	Route Filters, page 14-27	March 1, 2018
Maximum length of a calling search space	Calling Search Spaces, page 14-43	March 1, 2018
SIP route header	Routing of SIP Requests in Unified CM, page 14-48	March 1, 2018

Dial Plan Fundamentals

Developing an end-to-end enterprise dial plan requires a sound understanding of a number of concepts that are independent of specific products. This section provides an overview of those concepts, including:

- Endpoint Addressing, page 14-3
- Dialing Habits, page 14-6
- Dialing Domains, page 14-7
- Classes of Service, page 14-8
- Call Routing, page 14-8

Endpoint Addressing

Reachability of endpoints registered to a call processing agent, users, and applications is provided by addresses assigned to these addressable entities. In enterprise collaboration networks we differentiate between numeric addresses and alphanumeric uniform resource identifiers (URIs).

Numeric Addresses (Numbers)

Numeric addresses are represented by a sequence of digits. The call control agent does not assume, preclude, or require a special structure for numeric addresses. The decision on the structure of the numeric addresses to be used in the dial plan is part of the dial plan design process.

ITU recommendation E.164 defines the fundamental structure of numeric geographical addresses (phone numbers) to be used in the PSTN, as shown in Figure 14-1.

Figure 14-1 E.164 Format for Geographic Numbers

<table>
<tr><td rowspan="2"></td><td colspan="2">Maximum of 15-n digits (n = number of CC digits)</td></tr>
<tr><td>1 to 3 digits</td><td>Defined by National Numbering Plan</td><td>Defined by National Numbering Plan</td></tr>
<tr><td colspan="3">Maximum of 15 digits</td></tr>
</table>

The following definitions apply to Figure 14-1:

- CC = Country Code
- NSN = National Significant Number
- NDC = National Destination Code
- SN = Subscriber Number

According to ITU recommendation E.164, the maximum length of any phone number is 15 digits. The first part of a geographic E.164 number is the country code. Country codes are between one and three digits long (country code 1 and 7 are the only single-digit country codes). The remainder of a geographic E.164 number is the national significant number (NSN). The general structure of a NSN is that the first few digits represent a national destination code (NDC), or area code, and the last digits represent the

subscriber number. ITU recommendation E.164 does not define national numbering plans and thus does not prescribe the schema to be used for NSNs in specific countries. This is left to the national numbering plan authorities. A collection of documentation on various national numbering plans can be found at

https://www.itu.int/oth/T0202.aspx?parent=T0202

National numbering plans can be very different in structure. As an example, Table 14-2 compares the numbering plans used in the US and in Germany.

Table 14-2 Comparison of Numbering Plans in the US and Germany

Country Code	NSN Length	NDC Length	SN Length
1 (US)	10	3	7
49 (Germany)	3 to 13	2 to 5	4 to 11 (depending on area code)

ITU recommendation E.164 also mentions that a leading "+" should be used to indicate if an international prefix is required. Throughout this design guide we consistently use the term "E.164" to refer to E.164 numbers and "+E.164" to refer to E.164 numbers with a leading "+".

Using +E.164 numbers as numeric addresses has the benefit that these numbers by definition are unique and that it is very easy to map between +E.164 and any habitual dialing that might be required to be supported by an enterprise dial plan.

As an alternative to using unique numeric +E.164 addresses, numeric addresses following an enterprise numbering plan may also be used. Building an enterprise address plan or numbering scheme in multi-site deployments involves the definition of a typical hierarchical addressing structure with the following characteristics:

- Provides unique numeric addresses for all endpoints, users, and applications in the enterprise.
- Needs to be extensible so that the numbering scheme allows for adding new sites without having to redesign the whole numbering scheme, which would involve address reassignments for existing endpoints, users, and applications.

In a typical enterprise numbering plan, numerical addresses would consist of a site code and a site subscriber number. When designing an enterprise numbering plan, reserve enough digits for both the site code and the site subscriber number to make sure that additional sites can be added if required and enough subscriber numbers can be defined per site. Enterprise numbering plans typically are designed to be fixed length.

If or when dialing of addresses defined by an enterprise numbering plan needs to be supported directly as a dialing habit, typically a single-digit access code is selected and prefixed to the enterprise number. In that case a full enterprise numbering address would have three pieces: enterprise address access code (for example, 8), site code (for example, 496), and site subscriber number (for example, 9123); or for example, 8-496-9123.

The enterprise address access code in this case needs to be selected so that it does not cause overlap with any other dialing habit (see Dialing Habits, page 14-6).

To be able to uniquely identify an addressable entity, either all addresses have to be unique or they at least have to be unique within a defined sub-domain managed by the call processing agent. If two distinct entities need to be addressed using the same address, then the two identical addresses have to reside in disjunct addressing domains that are managed independently. With numeric addresses, this situation can occur if site-significant numeric addresses (for example, a four-digit extension) are used and two endpoints with the same site-significant address (same four-digit extension) in different sites need to be addressed by the same call control agent. Table 14-3 shows an example of this situation.

Table 14-3 Overlapping Numeric Addressing

+E.164 Number	Site (Sub-Domain)	Site DID Range	4-Digit DN (Address)
+49 690 773 3001	Frankfurt	+49 690 773 3XXX	3001
+1 408 555 3001	San Francisco	+1 408 555 3XXX	3001

In Table 14-3, two E.164 numbers result in the same site-specific four-digit directory number based on the respective site's DID range. This implies that the four-digit DNs cannot be used as numeric endpoint addresses directly.

Addresses following an enterprise numbering plan, also known as enterprise significant numbers (ESN), can be used to address destinations for which no PSTN numbers (E.164 numbers) exist. These destinations include:

- Lobby phones
- Regular endpoints for which no DIDs can be assigned by the provider
- Services (call pickup numbers, call park numbers, conferences, and so forth)

Alphanumeric Addresses

Alphanumeric addresses based on SIP URIs can also be used to address endpoints, users, and applications. A commonly used scheme for alphanumeric addresses is simplified SIP URIs of the form *user@host*, where the left-hand side (LHS, user portion) can be alphanumeric and the right-hand side (RHS, host portion) is a domain name. The following examples represent valid alphanumeric addresses based on SIP URIs:

- bob@example.org
- bob.home-office@example.org
- bob@de.eu.example.org
- bob.ex60@example.org
- bob.vmbox@example.org
- voicemail@de.eu.example.org

All of these URIs can serve as individual alphanumeric addresses for individual endpoints, users, and applications. From the addressing perspective, any hierarchy implied by using dot-separated identifiers (bob.ex60, de.eu.example.org) does not have any impact on the decision of whether any two URIs are considered to be equivalent.

According to RFC 3261, comparison of the LHS of SIP URIs has to be case-sensitive, while the RHS has to be compared case-insensitive. According to this standardized behavior, Alice@example.com and alice@example.com are not to be considered equivalent and thus represent distinct individual addresses. In reality, using addressing schemes for endpoints, users, and applications that rely on case sensitivity of the user portion is considered bad practice because it leads to increased troubleshooting complexity. Also keep in mind that RFC 2543 (the RFC specification preceding RFC 3261) explicitly defined that SIP URIs (host and user portion) are case-insensitive. Different behaviors regarding the case sensitivity of URI equivalence are common. To avoid problems, Cisco highly recommends always using only all lowercase URIs as alphanumeric addresses.

The URI lookup policy of Unified CM can be configured to be case-sensitive (default) or case-insensitive by configuring the enterprise parameter **URI Lookup Policy** accordingly.

Dialing Habits

Dialing destinations such as users, endpoints, and applications can be done using various formats. The numeric +E.164 address +496907739001, for example, could be dialed as any of the following:

- +496907739001
- 9011496907739001 from an enterprise extension in the US
- 011496907739001 from a land-line phone in the US
- 006907739001 from an enterprise extension in Germany
- 000496907739001 from an enterprise extension in Italy
- 9001 from a phone in the same office

These examples show that dial strings typically are interpreted in a context, and with the exception of dialing a +E.164 address directly, only the combination of dialed string and context provides proper identification of the intended destination address.

The term "dialing habit" is commonly used to refer to a given dialing behavior used to reach a given set of destinations. Examples of some “dialing habits” include:

- 9-0-1-1 plus E.164 for international destinations dialed from within the US
- 0-0 plus NSN for national calls in Germany
- 9-1 plus 10 digits for national calls in the US
- Four digits for intra-site calls in an office

A dialing habit is described by specifying both the format of the string to be dialed (dialing structure) and the destination address class to be reached. Examples of destination address classes typically used in enterprise dial plans include:

- On-net/intra-site
- On-net/inter-site
- Off-net/local
- Off-net/national
- Off-net/international
- Off-net/emergency
- Services (voicemail, pick-up, and so forth)

Identifying the dialing habits that need to be supported by the dial plan is one of the first steps when starting the design of an enterprise dial plan. It is essential to start the dial plan design with the full view of all dialing habits to be supported, because the dialing habits need to be defined so that there is no overlap between any two dialing habits to be supported for any given caller. Violating this rule will lead to bad user experience because the call control cannot deterministically differentiate between overlapping dialing habits by analyzing the dialed digits as they are dialed one-by-one. This ultimately leads to inter-digit timeouts.

With alphanumeric dialing we typically differentiate only between fully qualified addresses and non-fully qualified addresses. Fully qualified addresses contain the user and the host portion of a SIP URI, whereas a non-fully qualified alphanumeric address represents only the user portion of the address and the host portion needs to be derived from the dialing context of the calling party. For example, dialing "bob" would be equivalent to dialing "bob@example.com" if the dialing context of the calling party defines that "@example.com" should be appended to all non-fully qualified alphanumeric addresses.

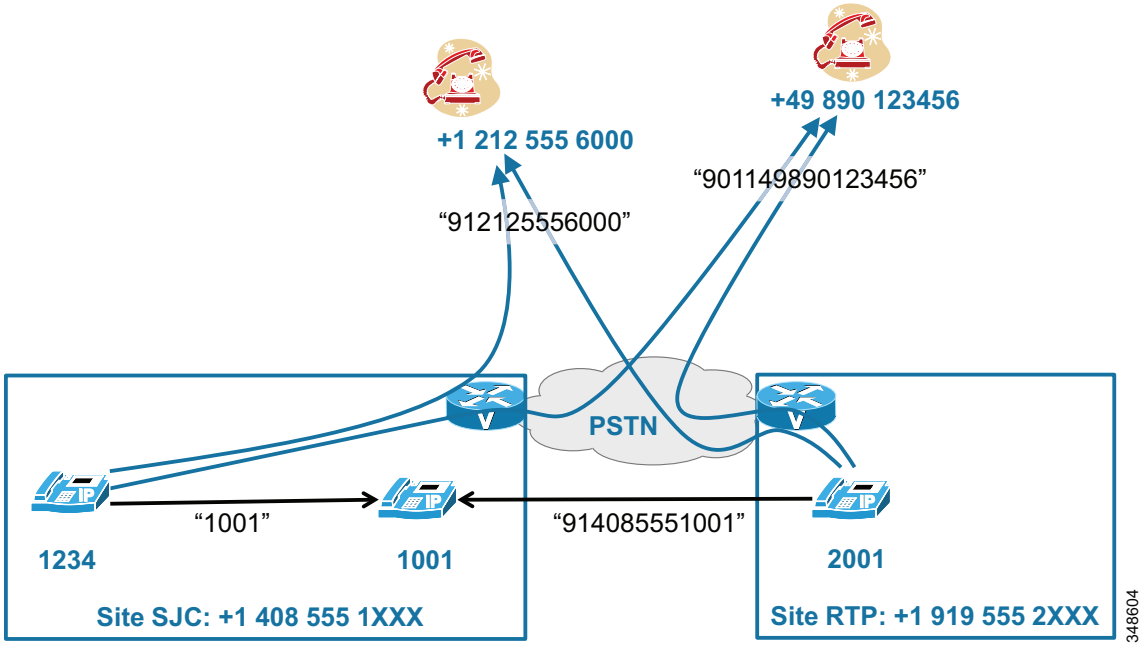
Dialing Domains

As described in the previous section, a given destination might be dialed using different strings by different users. A dialing domain specifies a group of users or devices sharing the same set of dialing habits (dialing the same strings to reach identical destinations). The concept of dialing domains is important because an enterprise dial plan has to implement the same treatment for each dialing domain. All users belonging to any given dialing domain share the same dialing treatment.

To identify dialing domains, it is important to take all dialing habits into consideration. Users in two sites in the US, even though they share the same PSTN dialing habits, would still belong to different dialing domains if we also take into account how on-net calls are placed. In a typical environment, an on-net intra-site call could be supported by dialing four digits, while an on-net inter-site call would be placed by using a dial string equivalent to the PSTN dialing habit (forced on-net would still keep the call on-net).

Figure 14-2 shows an example for this. Although endpoint 1234 in site SJC and endpoint 2001 in site RTP share the same dialing habit for national destinations (dialing 91212555600 to reach PSTN destination +1 212 555 6000) and international destinations (dialing 901149890123456 to reach PSTN destination +49 890 123456), the dialing habit to reach endpoint 1001 in site SJC is different for endpoints in RTP than for those in SJC: endpoint 1234 in site SJC would dial 1001 while endpoint 2001 in site RTP would need to dial 914085551001. In this example, endpoints in site RTP and site SJC would belong to different dialing domains.

Figure 14-2 Dialing Domains



Classes of Service

Not all users, applications, and endpoints in an enterprise are allowed to reach the same set of destinations. Reasons for restricting the set of reachable destinations include cost avoidance, security considerations, and privacy. As examples, not all users might be able to place international calls, the voicemail systems might not be able to call any PSTN destination to avoid toll fraud, and only a very limited set of users might be allowed to place direct calls to the CEO of a company. The term generally used to refer to any given set of a restrictions or class of allowed destinations is *class of service*, or CoS.

Requirements for cost-driven classes of service heavily depend on the phone tariffs and the cost structure associated with them. With voice services becoming cheaper (or being available for free), the trade-off between the increased complexity associated with maintaining more classes of service and the potential savings in call costs is changing. In certain cases, for example, it might not make sense any more to differentiate between local and national calling if both call types are billed exactly the same.

The definition of a class of service might also be based on time schedules. Access to certain destinations might be allowed only at certain times.

To reduce the complexity of an enterprise dial plan, Cisco recommends minimizing the number of differentiated classes of service. This can be achieved by either removing classes of service with little or no value (for example, differentiation between classes of service "national" and "local" even though national calls are essentially for free) or by combining (almost) equivalent classes of service into a single class of service.

Independent of restricting the access of certain users, devices, or applications to certain call types that incur costs, typically access to emergency services (911, 112, and so forth) has to be provided to all users at all times. Therefore, all classes of service have to allow access to emergency services at all times.

Call Routing

Routing calls involves several aspects:

- Identifying the dialing habit based on the structure of the dial string.
- Allowing/blocking the call based on the class of service of the calling entity.
- Applying modifications to the dial string.
- Applying modifications to the calling party identification.
- Selecting a route to the called destination, establishing the call, and presenting the identity of the parties involved in the expected format. Route selection also involves selecting alternate routes if the primary route is not available for some reason.

An end-to-end enterprise dial plan needs to consider all of these aspects and is not limited only to establishing a route between the calling and called entities.

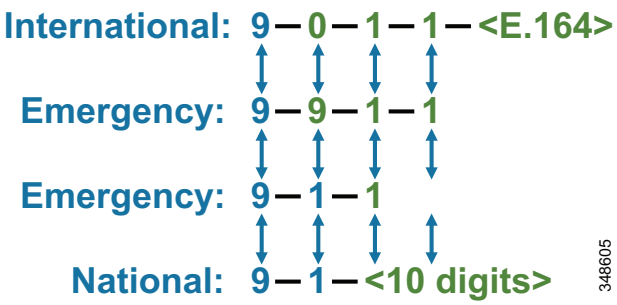
Identification of Dialing Habit and Avoiding Overlaps

The first step in the call routing process, after receiving the input from the calling entity, is to uniquely identify the dial habit used. In the case of alphanumeric dialing, this is a trivial task because in this case we typically only need to differentiate between fully qualified SIP URIs and non-fully qualified SIP URIs. This can easily be achieved through a simple lexical analysis of the dialed string.

The case of numeric dialing needs a little more attention, especially if the dialed digits are entered digit-by-digit. In this case the call control receives the destination from the calling entity digit-by-digit, and part of selecting the correct route to the destination is to determine the exact time when enough digits have been received and the call can be routed without having to wait for expiration of an inter-digit timeout.

Figure 14-3 shows some typical US dialing habits for PSTN and emergency calls. Although all of these dialing habits share the identical initial digit 9, international dialing and the first emergency dialing string can easily be distinguished as soon as the second digit (0 or 9) is dialed. As soon as the third digit is dialed, dialing 911 and dialing a national destination do not overlap any more because the North American Numbering Plan (NANP) does not allow any NPA codes (numbering plan area codes) starting with 1.

Figure 14-3 Typical US Dialing Habits for PSTN and Emergency Calls



Given the PSTN dialing habit in Figure 14-3. four-digit intra-site dialing for extensions starting with 9 must be avoided because this could potentially create partial overlap. For example, extension 9113 would overlap with emergency calling, and after receiving 911 the call control would have to wait to determine whether the caller is going to continue to dial 3 (extension 9113) or whether dialing actually was complete after 911. This would delay all emergency calls! Similarly, extensions such as 9140 would create overlaps with national PSTN calls, and calls to those extensions would be delayed.

To minimize overlaps, the first digit of a dialing habit can be defined as an access code uniquely identifying a class of destinations. The PSTN or trunk access code is a perfect example for this scheme. The most commonly used trunk access codes are 9 (US, UK, and others) and 0 (commonly used in most European countries).

As mentioned earlier, selecting non-overlapping dialing habits is key to avoiding bad user experience due to inter-digit time-outs. Typical overlaps seen in enterprise dial plans include:

- 10-digit dialing with abbreviated intra-site dialing (for example, four digits)

  NPA codes in the US can start with any digit other than 0 or 1, which means that the first few digits of 10-digit dialing would overlap with almost any abbreviated intra-site dialing.

- PSTN access code (such as 9 in the US) with abbreviated intra-site dialing

  A PSTN access code of 9 will overlap with all abbreviated intra-site dialing to extensions starting with 9.

- Abbreviated on-net inter-site dialing and abbreviated intra-site dialing

  The access code selected for the abbreviated on-net enterprise numbering plane might overlap with the range of intra-site dialing starting with the same digit. For example, using access code 8 for abbreviated on-net intra-site dialing prohibits the use of abbreviated intra-site dialing starting with 8.

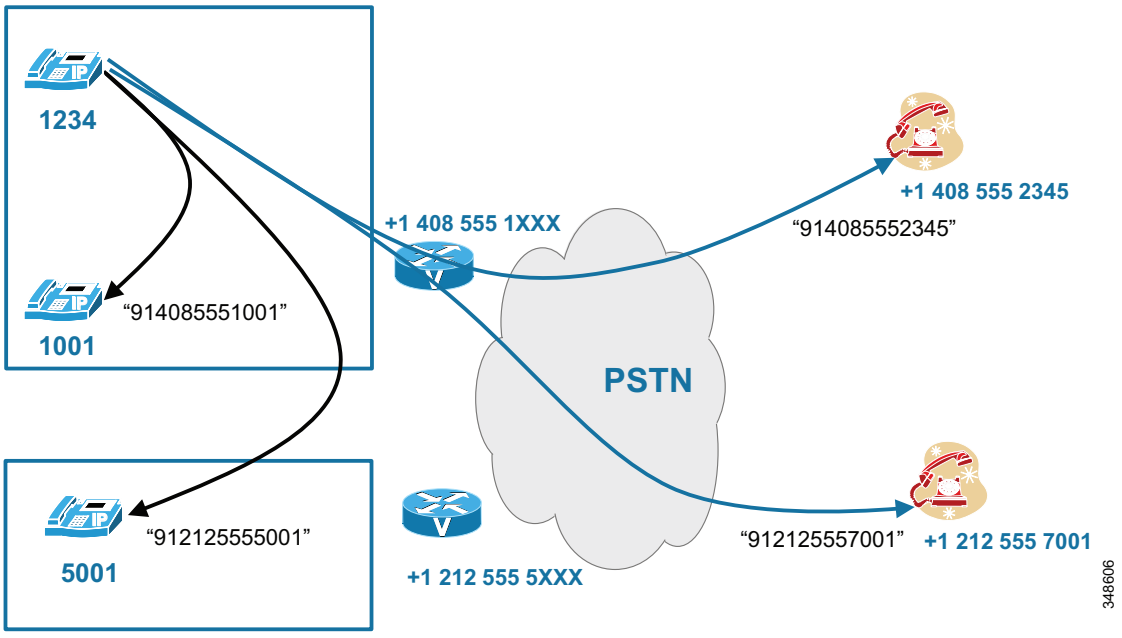
Access to features such as call park numbers and voicemail also requires mapping into the set of dialing habits defined. Dialing these features should typically require only a short dialing sequence. To achieve this, either the feature access codes can be mapped into the abbreviated intra-site dialing or a dedicated feature access code needs to be defined.

Forced On-Net Routing

It is not uncommon for the dialing habits for on-net/inter-site and off-net destinations to use the same dialing structure. In this case the call control decides whether the addressed endpoint, user, or application is on-net or off-net based on the dialed address, and will treat the call as on-net or off-net, respectively.

Figure 14-4 shows an example of this forced on-net routing. All four calls in this example are dialed as 91 plus 10 digits. But while the calls to +1 408 555 2345 and +1 212 555 7000 are really routed as off-net calls through the PSTN gateway, the other two calls are routed as on-net calls because the call control identifies the ultimate destinations as on-net destinations. Forced on-net routing clearly shows that the dialing habit used to dial a specific destination does not necessarily determine how a call is routed. In this example, some calls are routed as on-net calls even though the used PSTN dialing habit seems to indicate that an off-net destination is called.

Figure 14-4 Forced On-Net Routing



Forced on-net routing is especially important if dialing of +E.164 destinations from directories is implemented. In a normalized directory, all destinations are defined as +E.164 numbers, regardless of whether the person that the number is associated with is internal or external. In this case forced on-net routing is a mandatory requirement to avoid charges caused by internal calls routed through the PSTN.

Single Call Control Call Routing

If all endpoints are registered to a single call control, this call control has a full view of all known on-net destinations. When a user, endpoint, or application dials a destination using any of the defined dialing habits, the call control can easily determine whether the dialed destination is on-net or off-net. This might be based on the used dialing habit or on the normalized dialed address (see Forced On-Net Routing, page 14-10).

If the dialed destination is determined to be external, the call control then needs to select the correct external route to set up the call. If only one external (PSTN) connection exists, this is a trivial decision. If multiple external connections exist, the egress route selection can be based on any combination of the following factors:

- Call initiator
- Dialed destination
- Resource availability
- Resource prioritization

To be able to select an external connection based on the dialed destination, the dialed destination must be classified. As explained earlier, E.164 numbers have a hierarchical structure that implies some geographic association of numbers so that the egress connection selection could be based on a prefix-based hierarchical routing scheme that tries to select an egress point "closest" (in the sense of the geographic semantics of the E.164 number) to the destination. This behavior is called Tail End Hop Off (TEHO). When implementing Tail End Hop Off, local legal regulations have to be considered.

An interesting special case of TEHO exists if strange phone tariffs allow for cheaper national calls (for example, interstate) than local calls (in-state). In this case an egress point selection policy might be implemented that actually tries to avoid selecting an egress connection "too" close to the dialed destination. Decreasing phone charges make these kinds of routing schemes less and less useful.

In contrast to E.164 numbers, which have a clear hierarchical geographical structure with the most significant information on the left alphanumeric, SIP URIs allow addressing hierarchy in the host portion (RHS) of the URI. Depending on the domain name used as RHS, the addressing hierarchy of URIs does not necessarily allow for geographic mapping a URI to a location in the routing topology, especially if a flat URI scheme such as user@example.org is used. More interestingly, the most significant piece of a SIP URI is the right-most piece of the host portion (top level domain).

Multiple Call Control Call Routing

In larger enterprise networks, a number of call controls might be deployed. These independent call controls are interconnected using trunks. The possible topologies include hub and spoke, full mesh, and combinations of these. Any of these call controls will independently route calls offered by either endpoints, applications registered locally, or internal and external trunks.

In an environment like this, the on-net/off-net decision described in the previous section gets a little more complex. Before routing a call to an external connection, each call control needs to be sure that the dialed destination really is off-net. But by looking at only the locally registered addresses, the call control can actually get to only a reliable local/remote decision, and any destination classified as remote (not locally registered) can still be on-net but hosted on one of the other enterprise call controls deployed.

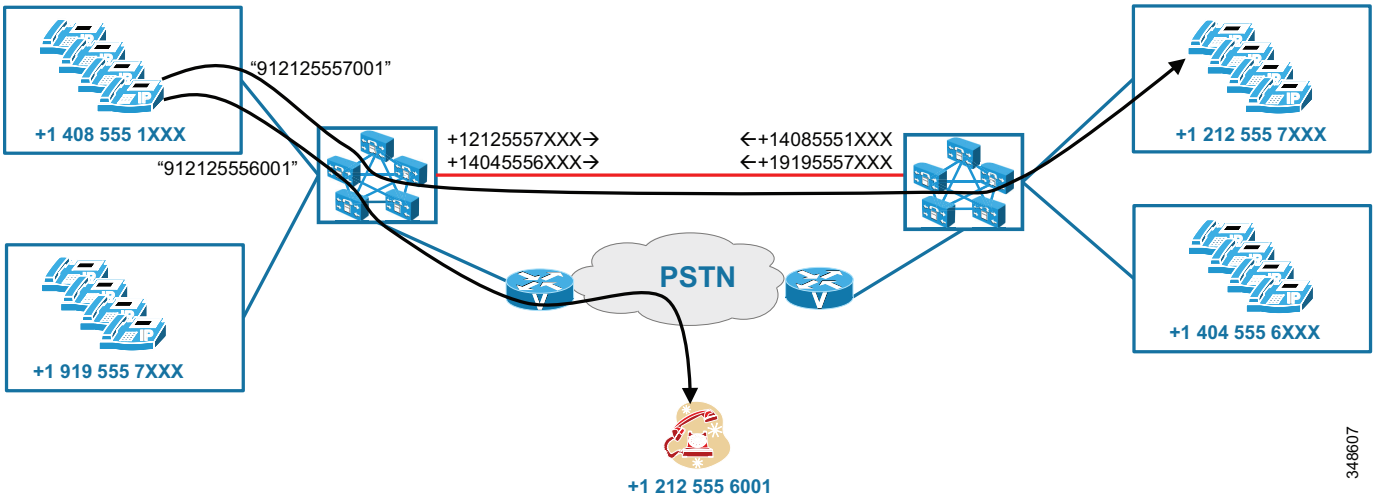
With numeric addressing and strict geographic assignment of endpoints to individual call controls, selecting the correct internal or external connection on any given call control comes down to implementing a routing scheme based on E.164 prefixes. This essentially is identical to the call routing process described for the single call control case, with the only difference being that some of the connections used for the prefix-based routing are not external connections (for example, to the PSTN)

but are internal connections to other enterprise call control entities. To make sure that only calls to remote on-net destinations are routed to the remote call control, the call routing decision needs to be based on the specific address ranges local to the remote call control.

Figure 14-5 shows why the prefix-based routing between independent call controls must be very specific. In this example, to enable the left call control to decide wether 912125556001 needs to be treated as an on-net call, the left call control has to have very specific numeric prefix routes for all numeric addresses served by the right call control entity.

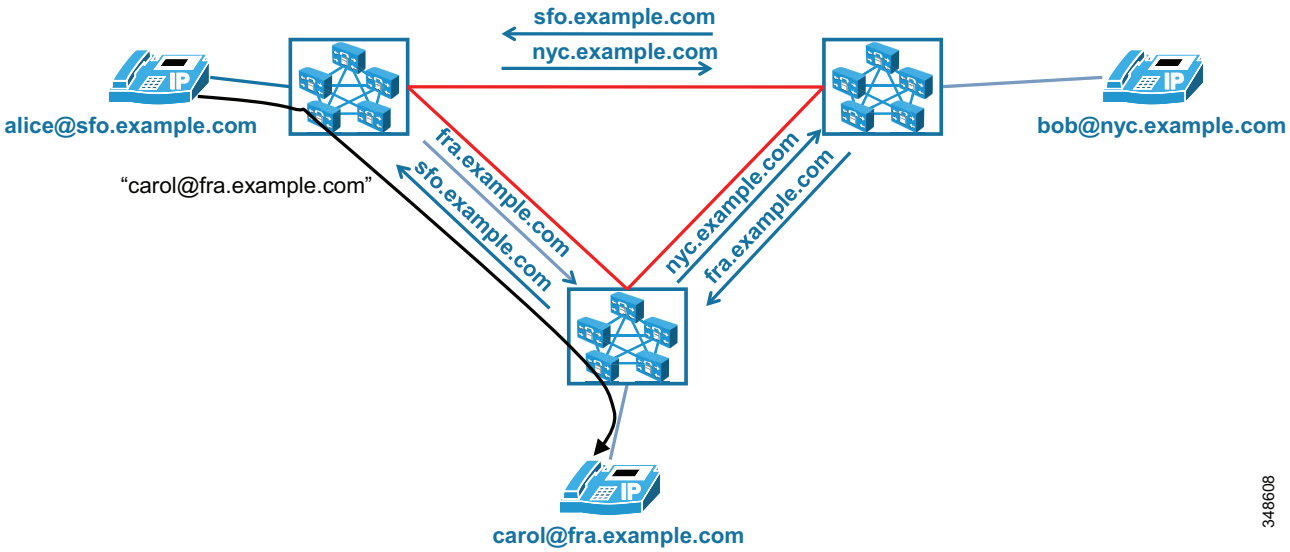
The maintenance of on-net prefix routes provisioned on each call control becomes more complex with an increasing number of call controls involved and sites and DID ranges to be considered. Dynamic learning of remote destinations helps to eliminate this complexity. Global Dial Plan Replication (GDPR) is one example of an architecture that allows call controls to automatically learn about destinations existing on remote call controls.

Figure 14-5 Prefix-Based Routing Between Call Controls



The equivalence of prefix-based routing of numeric addresses for alphanumeric URIs is to use a domain hierarchy and implement routing based on the host or domain portion of the URI. Figure 14-6 shows an example of hierarchical routing with alpha URIs. In this example all three independent call controls use a dedicated (sub) domain so that the on-net routing can easily be implemented based on this hierarchical domain structure.

Figure 14-6 Hierarchical Routing for Alpha URIs



In cases where the URI addressing scheme is not hierarchical, each call control has to have knowledge of all URIs hosted on remote call controls. Global Dial Plan Replication (GDPR) offers a mechanism for call controls to exchange information about URIs hosted on each call control to enable deterministic routing even with a flat URI naming scheme.

Dial Plan Elements

This section describes the dial plan elements available in these solution components:

- Cisco Unified Communications Manager, page 14-13
- Cisco TelePresence Video Communication Server, page 14-53

Cisco Unified Communications Manager

This section provides design and configuration guidelines for the following dial plan elements within Cisco Unified Communications Manager (Unified CM):

- Calling Party Transformations on IP Phones, page 14-14
- Support for + Dialing on the Phones, page 14-15
- User Input on SCCP Phones, page 14-15
- User Input on Type-A SIP Phones, page 14-16
- User Input on Type-B SIP Phones, page 14-18
- SIP Dial Rules, page 14-20
- Call Routing in Unified CM, page 14-22
- Translation Patterns, page 14-24
- Calling Privileges in Unified CM, page 14-41

User Interface on IP Phones

Note Different types of IP telephones accept keypad input and present visual information in different ways. For purposes of this chapter only, we define the following phone types:

- **Type-A phones** — Include the Cisco Unified IP Phone 7905, 7912, 7940, and 7960.
- **Type-B phones** — Include the Cisco Unified IP Phone 6901, 6911, 6921, 6941, 6945, 6961, 7906, 7911, 7921, 7925, 7931, 7941, 7942, 7945, 7961, 7962, 7965, 7970, 7971, 7975, 8961, 9951, 9971, and newer phones.

Calling Party Transformations on IP Phones

Calling Party Transformation Patterns allow the system to adapt the calling party numbers to different formats. The most typical use is to adapt from globalized to localized calling party numbers and vice versa.

The transformation pattern consists of a numerical representation of the calling party number to be matched. The syntax used is the same as that of other patterns such as route patterns, called party transformation patterns, directory numbers, and so forth.

The transformation operators available on a transformation pattern include discard digit instructions (for example, pre-dot), a calling party transformation mask, prefix digits, and an option to apply the calling party's external phone number mask. In addition, the calling party presentation indicator can be set (either Default, Allowed, or Restricted).

Partitions and calling search spaces control which calling party transformation patterns are applied to which phones. Phones can use either the device pool's calling party transformation calling search space (CSS) or the device's own calling party transformation CSS, in reverse order of precedence.

On IP phones, calling party transformations can be configured for calls originating from the phone and for calls terminated on the phone:

- For calls originating from phones where the configured directory numbers are not in a globalized (+E.164) form, the inbound call's calling party transformation CSS can be used to define the appropriate globalization. This CSS can be found on the phone configuration page in the Number Presentation Transformation section or in the Phone Settings section on the device pool configuration page under **Caller ID For Calls From This Phone**.
- For calls terminating on phones, the outbound calls' calling party transformation CSS can be used to define the localization scheme to be applied to calling party numbers. This CSS can be found on the phone configuration page in the Number Presentation Transformation section under **Remote Number** or on the device pool configuration page under **Device Mobility Related Information as Calling Party Transformation CSS**.

For phones, outbound or remote number calling party transformations affect the number displayed while the phone is ringing.

The outbound call's calling party transformation CSS (also referred to as localization or remote number calling party transformation CSS) can also be used to localize remote connected party information. To enable this, the advanced service parameter **Apply Transformations On Remote Number** must be enabled.

Being able to provide localized connected party information to phones enables consistent remote party information display on IP phones even if mid-call features are invoked.

Support for + Dialing on the Phones

On Type-A phones, it is not possible to dial a + sign using the keypad. On Type-B phones it is possible to dial a + sign by pressing and holding either the 0 key (Cisco Unified IP Phones 7921 and 7925) or the * key (all other phone models). On Cisco Unified Personal Communicator endpoints, the + sign may be typed in by the user using the computer's keyboard or entered as part of the input string when using a click-to-dial function of the endpoint.

On Type-A phones, there is no support to dial a +. The + sign can be displayed as part of the calling party information for incoming calls and in directories, but the phone will strip the + sign when the entry is dialed from the missed calls directory. To avoid misdialed calls, Type-A phones put the transformed number in the missed calls directory, and the callback also uses the transformed number. The transformed number has to be in the form of a dialing habit supported by the dial plan to avoid misdialed calls from directories.

On some endpoints, incoming calls can present a calling party number with + included as part of the number. When a call is offered to a phone, the number shown on the ringing phone is processed by any configured calling party number transformation patterns. The missed and received calls directories can hold both the original pre-transformation number and the transformed number. On some endpoints the number displayed in the list is the transformed number, and the pre-transformation number is visible only when looking at the details of an entry. The number dialed from the directory on some endpoints is the original pre-transformation number, allowing the one-touch dialing of previously received calls featuring the + sign as part of the calling number as long as the dial plan supports + dialing. On other endpoints the number dialed from the directory is the transformed number. To allow for one-touch dialing, this number needs to be in the format of a dialing habit supported by the dial plan.

Example 14-1 Calling Party Number with + Dialing

A Type-B phone in New York receives a call from +1 408 526 4000. Calling party transformation patterns are placed in the calling party transformation CSS in the phone's device pool. One of the patterns is configured as: \+1.!, strip pre-dot.

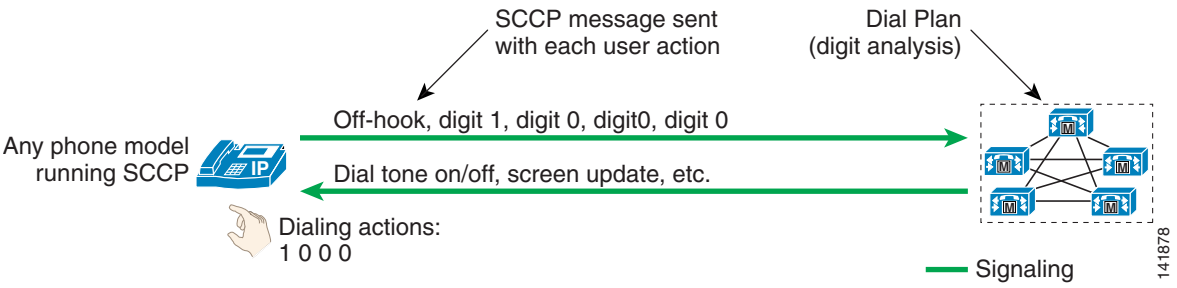
As the call rings in, the called phone displays an incoming calling party number of 4085264000. After the call is answered and released, the received-calls directory displays the last call as 408 526 4000, but the number called when the user initiated the callback from the directory entry is +1 408 526 4000.

User Input on SCCP Phones

IP phones using SCCP report every single user input event to Unified CM immediately. For instance, as soon as the user goes off-hook, a signaling message is sent from the phone to the Unified CM server with which it is registered. The phone can be considered to be a terminal, where all decisions resulting from the user input are made by Unified CM according to the configured dial plan.

As other user events are detected by the phone, they are relayed to Unified CM individually. A user who goes off-hook and then dials 1000 would trigger five individual signaling events from the phone to Unified CM. All the resulting feedback provided to the user, such as screen messages, playing dial tone, secondary dial tone, ring back, reorder, and so forth, are commands issued by Unified CM to the phone in response to the dial plan configuration. (See Figure 14-7.)

Figure 14-7 User Input and Feedback for SCCP Phones



It is neither required nor possible to configure dial plan information on IP phones running SCCP. All dial plan functionality is contained in the Unified CM cluster, including the recognition of dialing patterns as user input is collected.

If the user dials a pattern that is denied by Unified CM, reorder tone is played to the user as soon as that pattern becomes the best match in Unified CM's digit analysis. For instance, if all calls to the pay-per-minute Numbering Plan Area (or area code) 976 are denied, reorder tone would be sent to the user's phone as soon as the user dials 91976.

User Input on Type-A SIP Phones

Type-A phones differ somewhat from Type-B phones in their behavior, and Type-A phones do not offer support for Key Press Markup Language (KPML) as do Type-B phones. (See User Input on Type-B SIP Phones, page 14-18.)

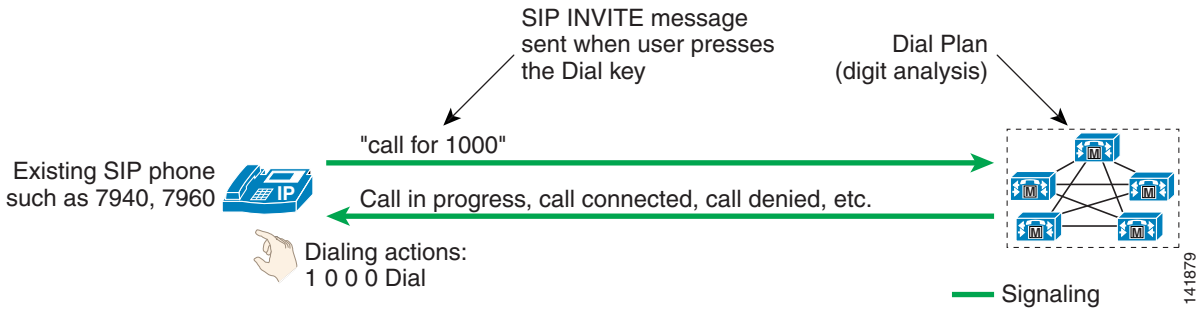
Type-A IP phones using SIP offer two distinct modes of operation:

- No SIP Dial Rules Configured on the Phone, page 14-16
- SIP Dial Rules Configured on the Phone, page 14-17

No SIP Dial Rules Configured on the Phone

Figure 14-8 illustrates the behavior of a SIP Type-A phone with no dial plan rules configured on the phone. In this mode of operation, the phone accumulates all user input events until the user presses either the # key or the Dial softkey. This function is similar to the "send" button used on many mobile phones. For example, a user making a call to extension 1000 would have to press 1, 0, 0, and 0 followed by the Dial softkey or the # key. The phone would then send a SIP INVITE message to Unified CM to indicate that a call to extension 1000 is requested. As the call reaches Unified CM, it is subjected to all the class-of-service and call routing logic implemented in Unified CM's dial plan.

Figure 14-8 User Input and Feedback for Type-A SIP Phones with No Dial Rules Configured



If the user dials digits but then does not press the Dial softkey or the # key, the phone will wait for inter-digit timeout (15 seconds by default) before sending a SIP INVITE message to Unified CM. For the example in Figure 14-8, dialing 1, 0, 0, 0 and waiting for inter-digit timeout would result in the phone placing a call to extension 1000 after 15 seconds.

Note If the user presses the Redial softkey, the action is immediate; the user does not have to press the Dial key or wait for inter-digit timeout.

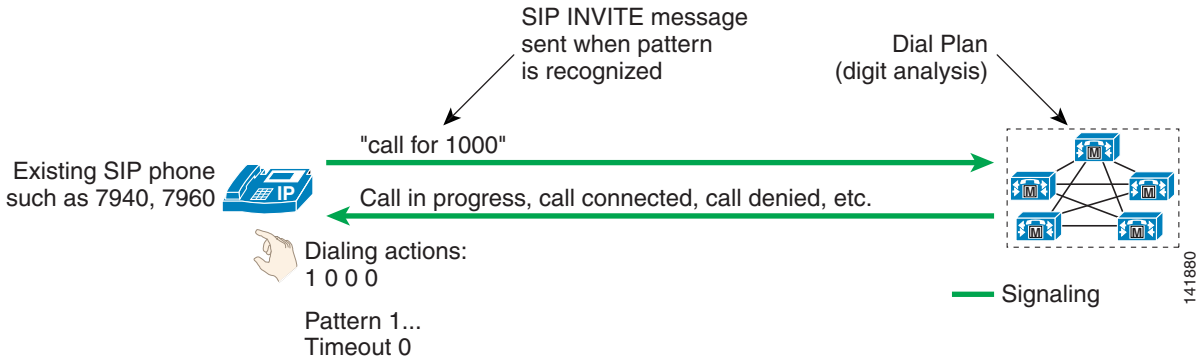
If the user dials a pattern that is denied by Unified CM, the user must enter the entire pattern and press the Dial key, and the INVITE message must be sent to Unified CM, before any indication that the call is rejected (reorder tone) is sent to the caller. For instance, if all calls to NPA 976 are denied, the user would have to dial 919765551234 and press Dial before the reorder tone would be played.

SIP Dial Rules Configured on the Phone

SIP dial rules enable the phone to recognize patterns dialed by users. Once the recognition has occurred, the sending of the SIP INVITE message to Unified CM is automated and does not require the user to press the Dial key or wait for the inter-digit timeout. (For more details, see SIP Dial Rules, page 14-20.)

For example, if a branch location of the enterprise requires that calls between phones within the same branch be dialed as four-digit extensions, the phone could be configured to recognize the four-digit patterns so that the user is not required to press the Dial key or wait for the inter-digit timeout. (See Figure 14-9.)

Figure 14-9 User Input and Feedback for Type-A SIP Phones with Dial Rules Configured



In Figure 14-9, the phone is configured to recognize all four-digit patterns beginning with 1 and has an associated timeout value of 0. All user input actions matching the pattern will trigger the sending of the SIP INVITE message to Unified CM immediately, without requiring the user to press the Dial key.

Type-A phones using SIP dial rules offer a way to dial patterns not explicitly configured on the phone. If a dialed pattern does not match a SIP dial rule, the user can press the Dial key or wait for inter-digit timeout.

If a particular pattern is recognized by the phone but blocked by Unified CM, the user must dial the entire dial string before receiving an indication that the call is rejected by the system. For instance, if a SIP dial rule is configured on the phone to recognize calls dialed in the form 919765551234 but such calls are blocked by the Unified CM dial plan, the user will receive reorder tone at the end of dialing (after pressing the final 4 key).

User Input on Type-B SIP Phones

Type-B phones differ somewhat from Type-A phones in their behavior, and Type-B phones offer support for Key Press Markup Language (KPML) but Type-A phones do not. (See User Input on Type-A SIP Phones, page 14-16.)

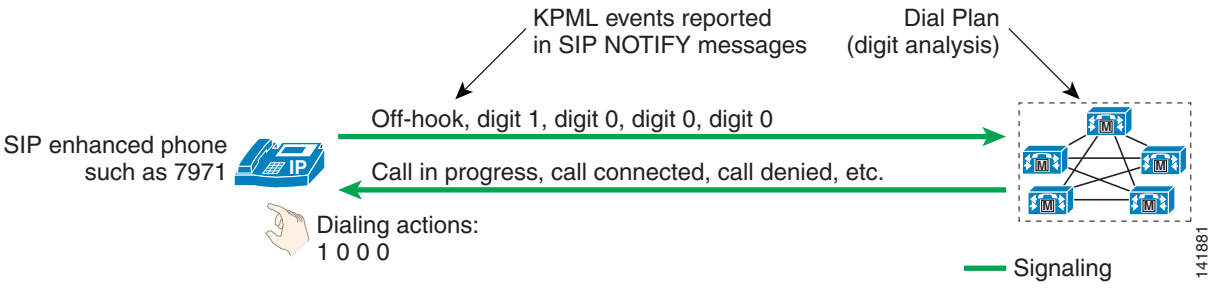
Type-B IP phones running SIP offer two distinct modes of operation:

- No SIP Dial Rules Configured on the Phone, page 14-18
- SIP Dial Rules Configured on the Phone, page 14-19

No SIP Dial Rules Configured on the Phone

Type-B IP telephones offer functionality based on the Key Press Markup Language (KPML) to report user key presses. Each one of the user input events will generate its own KPML-based message to Unified CM. From the standpoint of relaying each user action immediately to Unified CM, this mode of operation is very similar to that of phones running SCCP. (See Figure 14-10.)

Figure 14-10 User Input and Feedback for Type-B SIP Phones with No Dial Rules Configured



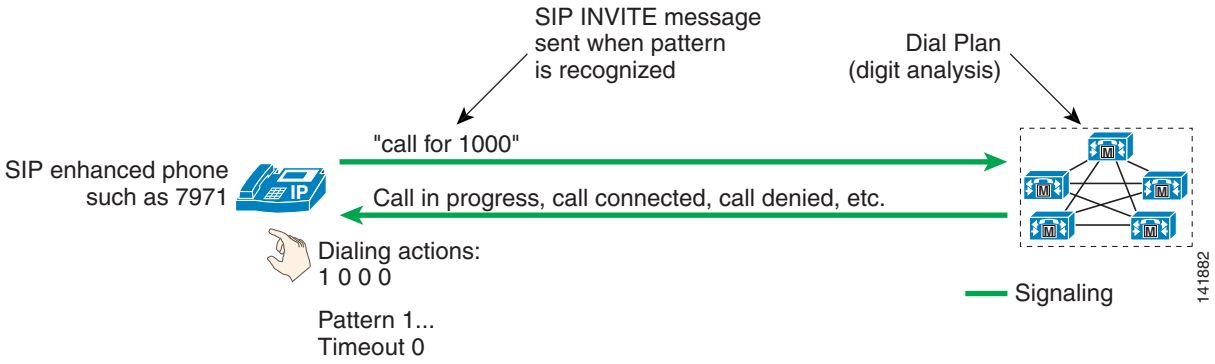
Every user key press triggers a SIP NOTIFY message to Unified CM to report a KPML event corresponding to the key pressed by the user. This messaging enables Unified CM's digit analysis to recognize partial patterns as they are composed by the user and to provide the appropriate feedback, such as immediate reorder tone if an invalid number is being dialed.

In contrast to Type-A IP phones running SIP without dial rules, Type-B SIP phones have no Dial key to indicate the end of user input. In Figure 14-10, a user dialing 1000 would be provided call progress indication (either ringback tone or reorder tone) after dialing the last 0 and without having to press the Dial key. This behavior is consistent with the user interface on phones running the SCCP protocol.

SIP Dial Rules Configured on the Phone

Type-B IP phones can be configured with SIP dial rules so that dialed pattern recognition is accomplished by the phone. (See Figure 14-11.)

Figure 14-11 User Input and Feedback for Type-B SIP Phones with Dial Rules Configured



In Figure 14-11, the phone is configured to recognize all four-digit patterns beginning with 1, and it has an associated timeout value of 0. All user input actions matching these criteria will trigger the sending of a SIP INVITE message to Unified CM.

Note As soon as SIP dial rules are implemented on Type-B IP phones, KPML-based dialing is disabled. If a user dials a string of digits that do not match a SIP dial rule, none of the individual digit events will be relayed to Unified CM. Instead, the entire dialed string will be sent en-bloc to Unified CM once the dialing is complete (that is, once inter-digit timeout has occurred).

Type-B phones using SIP dial rules offer only one way to dial patterns not explicitly configured on the phone. If a dialed pattern does not match a SIP dial rule, the user has to wait for inter-digit timeout before the SIP NOTIFY message is sent to Unified CM. Unlike Type-A IP phones, Type-B IP phones do not have a Dial key to indicate the end of dialing, except when on-hook dialing is used. In the latter case, the user can press the “dial” key at any time to trigger the sending of all dialed digits to Unified CM.

Note When a Type-B phone registers with the SRST router, the configured SIP dial rules are disabled.

If a particular pattern is recognized by the phone but blocked by Unified CM, the user must dial the entire dial string before receiving an indication that the call is rejected by the system. For instance, if a SIP dial rule is configured on the phone to recognize calls dialed in the form 919765551234 but such calls are blocked by the Unified CM dial plan, the user will receive reorder tone at the end of dialing (after pressing the 4 key).

SIP Dial Rules

Cisco Unified CM offers SIP dial rule functionality to allow phones to perform pattern recognition as user input is collected. For example, a phone can be configured to recognize the well established pattern 911 and to send a message to Unified CM to initiate an emergency call immediately, while at the same time allowing the user to enter patterns of variable length for international numbers.

It is important to note that pattern recognition configuration on the phone through the use of SIP dial rules does not supersede the Class of Service and Route Plan configurations of Unified CM. A phone might very well be configured to recognize long-distance patterns while Unified CM is configured to block such calls because the phone is assigned a class of service allowing only local calls.

There are two types of SIP dial rules, based on the phone model on which they will be deployed:

- 7905_7912 (Used for Cisco Unified IP Phones 7905 and 7912)
- 7940_7960_OTHER (Used for all other IP phone models))

There are four basic Dial Parameters that can be used as part of a dial rule:

- Pattern

  This parameter is the actual numerical representation of the pattern. It can contain digits, wildcards, or instructions to play secondary dial tone. The following table provides a list of values and their effect for the two types of dial rules.

	Dial Rule Type	
Pattern	**7905_7912**	**7940_7960_OTHER**
Digits 0 through 9	Corresponding digit	Corresponding digit
.	Matches any digit (0 through 9)	Matches any character (0 though 9, *, #)
-	Indication that more digits can be entered. Must be at the end of an individual rule.	n/a
#	Input termination key. Place the > character in the dial rule to indicate the character position after which the # key will be recognized as input termination. For instance, in 9>#..., the # character would be recognized any time after 9 has been pressed.	n/a
t*n*	Indicates a timeout value of *n* seconds. For example, 1…t3 would match 1000 and trigger the sending of an invite to Unified CM after 3 seconds.	n/a
r*n*	Repeats the last character *n* times. For example, 1.r3 is equivalent to 1….	n/a
S	When a pattern contains the modifier S, all other dial rules after this pattern are ignored. S effectively causes rule matching to cease.	n/a

Pattern	Dial Rule Type	
	7905_7912	7940_7960_OTHER
*	Input termination key. Place the > character in the dial rule to indicate the character position after which the * key will be recognized as input termination.	Matches one or more characters. For instance, pattern 1* would match 10, 112, 123456, and so forth.
,	n/a	Cause the phone to play secondary dial tone. For instance, 8,…. would cause the user to hear secondary dial tone after 8 is pressed.

- IButton

  This parameter specifies the button to which the dial pattern applies. If the user is initiating a call on line button 1, only the dial patterns specified for Button 1 apply. If this optional parameter is not configured, the dial pattern applies to all lines on the phone. This parameter applies only to the Cisco SIP IP Phones 7940, 7941, 7942, 7945, 7960, 7961, 7962, 7965, 7970, 7971, and 7975. The button number corresponds to the order of the buttons on the side of the screen, from top to bottom, with 1 being on top button.

- Timeout

  This parameter specifies the time, in seconds, before the system times out and dials the number as entered by the user. To have the number dialed immediately, specify 0. This parameter is available only for 7940_7960_OTHER dial rules. If this parameter is omitted, the phone's default inter-digit timeout value is used (default of 10 seconds).

- User

  This parameter represents the tag that automatically gets added to the dialed number. Valid values include **IP** (when SIP Call Agents other than Unified CM are deployed) and **Phone**. This parameter is available only for 7940_7960_OTHER dial rules. This parameter is optional, and it should be omitted in deployments where Unified CM is the only call agent. Keep in mind that a user=phone tag in a SIP request sent to Unified CM will force Unified CM to treat the SIP URI as a numeric URI. A SIP URI in the form of alice@cisco.com;user=phone will never be routed successfully because the user=phone tag forces numeric treatment and alice will not match any numeric pattern provisioned in Unified CM.

Note The Cisco Unified IP Phone 7905 and 7912 choose patterns in the order in which they have been created in the SIP dial rules, whereas all the other phone models choose the pattern offering the longest match. The following table shows which pattern would be chosen if a user dialed 95551212.

SIP Dial Rules	7905_7912	7940_7960_OTHER
…….. 9…….	Chooses first matching pattern: ……..	Chooses longest matching pattern: 9…….

Call Routing in Unified CM

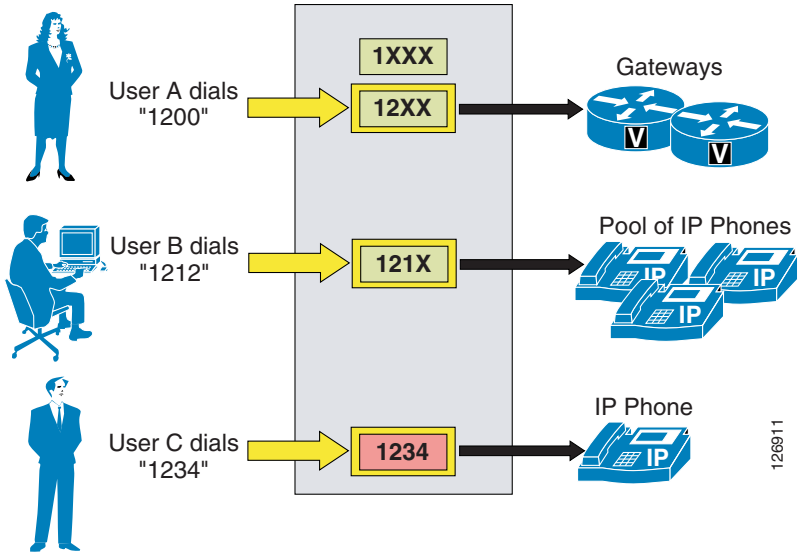
All numeric dialing destinations and directory URIs configured in Unified CM are added to its internal call routing table as patterns. These destinations include IP phone lines, voicemail ports, route patterns, translation patterns, and CTI route points. Unified CM uses two distinct routing tables for numeric dialing destinations and directory URIs.

When a directory URI is dialed, Unified CM uses full-match logic to find a match among the configured directory URIs in the directory URI routing table. The **URI Lookup Policy** enterprise service parameter setting determines whether the full-match logic for the user portion (left-hand side) of the URI uses case-sensitive or case-insensitive matching. Case-sensitive matching is the default. When a number is dialed, Unified CM uses best-match logic to select which pattern to match from among all the patterns in its numeric call routing table. In practice, when multiple potentially matching numeric patterns are present, the destination pattern is chosen based on the following criteria:

- It matches the dialed string, and
- Among all the potentially matching patterns, it matches the fewest strings other than the dialed string.

For example, consider the case shown in Figure 14-12, where the call routing table includes the patterns 1XXX, 12XX, and 1234.

Figure 14-12 Unified CM Call Routing Logic Example



When user A dials the string 1200, Unified CM compares it with the patterns in its call routing table. In this case, there are two potentially matching patterns, 1XXX and 12XX. Both of them match the dialed string, but 1XXX matches a total of 1000 strings (from 1000 to 1999) while 12XX matches only 100 strings (from 1200 to 1299). Therefore, 12XX is selected as the destination of this call.

When user B dials the string 1212, there are three potentially matching patterns, 1XXX, 12XX and 121X. As mentioned above, 1XXX matches 1000 strings and 12XX matches 100 strings. However, 121X matches only 10 strings; therefore it is selected as the destination of the call.

When user C dials the string 1234, there are three potentially matching patterns, 1XXX, 12XX, and 1234. As mentioned above, 1XXX matches 1000 strings and 12XX matches 100 strings. However, 1234 matches only a single string (the dialed string); therefore it is selected as the destination of this call.

When determining the number of matched strings for a variable-length pattern, Unified CM takes into account only the number of matched strings that are equal in length to the number of digits dialed. Assuming a user dials 1311 and we have patterns 1XXX, 1[2-3]XX, and 13!, the following table shows the number of matched strings of these potentially matching patterns.

Pattern	Number of Matched Strings	Possible Strings Matched
1XXX	1000	1000 to 1999
1[2-3]XX	200	1200 to 1299; 1300 to 1399
13!	100	1300 to 1399; only four-digit strings counted, based on the number of digits dialed

In this example the variable-length pattern 13! is selected as the best match.

Note Whenever a directory number (DN) is configured in Cisco Unified CM, it is placed in the call routing table, regardless of whether the respective device (for example, an IP phone) is registered or not. An implication of this behavior is that it is not possible to rely on secondary, identical patterns to provide failover capabilities to applications when the device (and hence the primary pattern) is unregistered. Because the primary pattern is permanently in the call routing table, the secondary pattern will never be matched.

Support for + Sign in Patterns

The + sign can be used in all patterns within Unified CM, including route patterns, translations patterns, and directory numbers. To use + in its literal sense, precede it with the escape character \ to differentiate it from the regular expression operator +, which means one or more instances of the preceding character. For example:

- \+14085264000 means +14085264000
- 2+ means 2, or 22, or 222, and so forth

This enables seamless implementation of +E.164 dial plans in Unified CM.

Directory URIs

All endpoints registered with Unified CM are provisioned with one or more numeric (possibly including a leading +) directory numbers. Up to five directory URIs can be associated with each directory number. This association can be created by explicitly associating directory URIs to directory numbers. If a directory URI is configured for an end user, this directory URI will be automatically associated with the primary extension of that end user as soon as the primary extension gets defined for that end user. All automatically associated directory URIs are created in the partition **Directory URI**, while manually configured directory URIs can be in any partition. Manually configured directory URIs can reside in the same partition as the directory number they are associated with, but do not have to. Directory URIs have to be unique per partition.

Exactly one of the directory URIs associated with a directory number has to be marked as the primary directory URI of that directory number. If a user directory URI gets associated automatically with the primary extension of that user, then this directory URI will also automatically be the primary directory URI for that directory number. If no directory URI is associated automatically, then one of the

configured directory URIs has to be selected as the primary directory URI. The purpose of the primary directory URI is that this directory URI will be used as the calling identity directory URI for calls originating from the respective directory number.

The possible association of directory URIs with any directory number allows callers to reach any directory number by dialing the associated directory URI. The called directory number can be on any device registered to Unified CM using any protocol. Similarly, Unified CM can deliver a directory URI caller ID for any call from any directory number as long as a directory URI is associated with the calling directory number.

To enable intercluster routing of directory URIs, Unified CM can be provisioned to exchange directory URI catalogs with other clusters through the Intercluster Lookup Service (ILS). Each cluster configured to exchange directory URI catalogs with other clusters advertises all locally configured directory URIs in a single directory URI catalog together with a location attribute, the SIP route string. This location attribute in multi-cluster environments is used to direct calls for directory URIs to the correct cluster when the host portion of the directory URI cannot be used to deterministically route the SIP request. This, for example, is the case when a flat URI scheme such as *<user>@example.com* is use. The host portion "example.com" does not uniquely identify the remote Unified CM cluster that hosts a given URI.

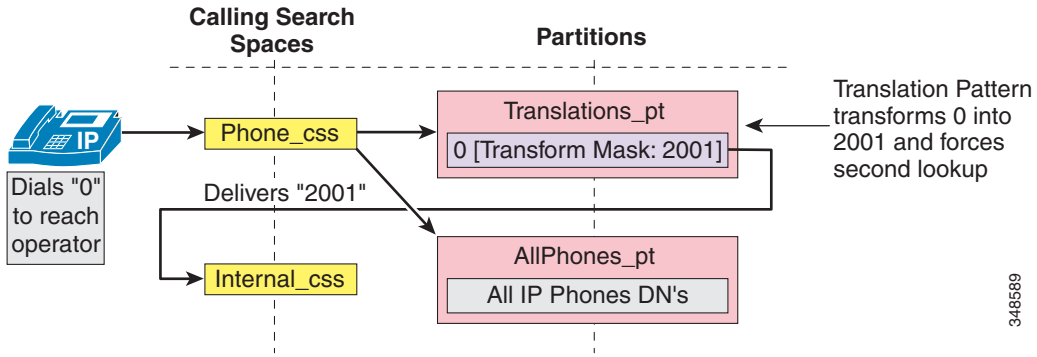
For details of how calls to directory URIs learned from remote clusters are routed, see the section on Routing of SIP Requests in Unified CM, page 14-48.

Translation Patterns

Translation patterns are one of the most powerful tools in Unified CM to manipulate digits for any type of call. They follow the same general rules and use the same wildcards as route patterns. As with route patterns, you assign a translation pattern to a partition. However, when the dialed digits match the translation pattern, Unified CM does not route the call to an outside entity such as a gateway; instead, it performs the translation first and then routes the call again, this time using the calling search space configured within the translation pattern.

Translation patterns can be used for a variety of applications, as shown by the example in Figure 14-13.

Figure 14-13 Application Example for Translation Patterns



In this example, the administrator wishes to provide users with an operator service that is reached by dialing 0, while also maintaining a fixed-length internal numbering plan. The IP phones are configured with the Phone_css calling search space, which contains the Translations_pt partition (among others). A translation pattern 0 is defined in this partition, and the configured Called Party Transform Mask instructs Unified CM to replace the dialed string (0) with the new string 2001, which corresponds to the

DN of the operator phone. A second lookup (of 2001 this time) is forced through the call routing engine, using the Internal_css calling search space, and the call can now be extended to the real operator DN of 2001, which resides in the AllPhones_pt partition.

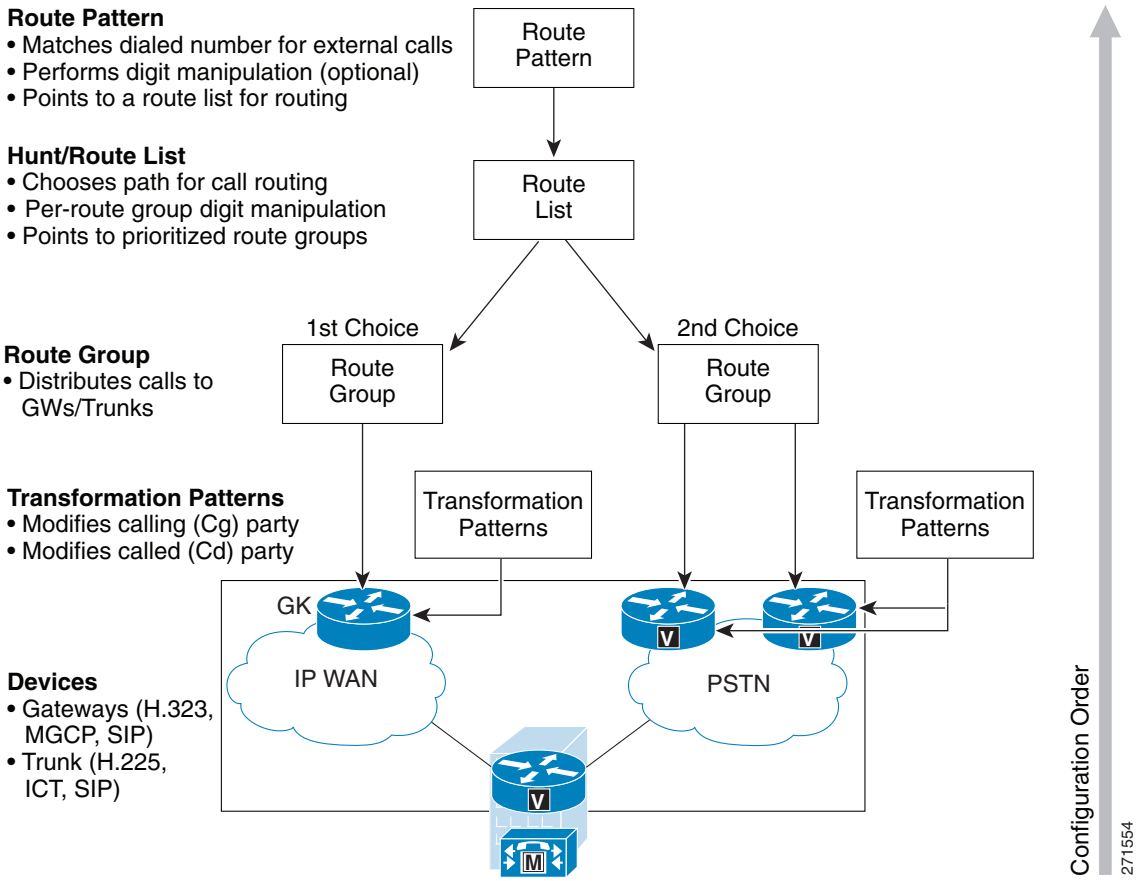
Note When a dialed number is manipulated using a translation pattern, the translated number is recorded in the call detail record (CDR). However, when the digit manipulation occurs within a route list, the CDR will show the originally dialed number, not the translated one. The Placed Calls directory on the IP phone always shows the string as it was dialed by the user.

The general use case for translation patterns is to create a mapping from a certain dial string format to a string to be matched by other dial plan elements. This mapping implements overlay dialing habits on top of the "native" dialing habits created by other patterns such as route patterns and directory numbers. Typically for the secondary lookup, translation patterns that implement a dialing normalization should simply use the calling search space that activates the translation pattern. This behavior, referred to as CSS Inheritance, is selected by the option **Use Originator's Calling Search Space** on the translation pattern. Enabling this option allows reuse of dialing normalization translation patterns for different classes of service, each defined by a different calling search space.

External Routes in Unified CM

Unified CM automatically "knows" how to route calls to internal destinations within the same cluster. For external destinations such as PSTN gateways, SIP trunks, or other Unified CM clusters, you have to use the external route construct (described in the following sections) to configure explicit routing. This construct is based upon a three-tiered architecture that allows for multiple layers of call routing as well as digit manipulation. Unified CM searches for a configured route pattern that matches the external dialed string and uses it to select a corresponding route list, which is a prioritized list of the available paths for the call. These paths are known as route groups and are very similar to trunk groups in traditional PBX terminology. Figure 14-14 depicts the three-tiered architecture of the Unified CM external route construct.

Figure 14-14 External Route Pattern Architecture



The following sections describe the individual elements of the external route construct in Unified CM:

- Route Patterns, page 14-26
- Route Lists, page 14-29
- Route Groups, page 14-30
- Route Group Devices, page 14-30

Route Patterns

Route patterns are strings of digits and wildcards, such as 9.[2-9]XXXXXX, configured in Unified CM to route calls to external entities. The route pattern can point directly to a gateway for routing calls or point to a route list, which in turn points to a route group and finally to a gateway.

Cisco strongly recommends that you use the complete route pattern, route list, and route group construct because it provides the greatest flexibility for call routing, digit manipulation, route redundancy, and future dial plan growth.

The @ Wildcard

- The @ wildcard is a special macro function that expands into a series of patterns representing the entire national numbering plan for a certain country. For example, configuring a single unfiltered route pattern such as 9.@ with the North American Numbering Plan really adds 166 individual route patterns to the Unified CM internal dial plan database.

- It is possible to configure Unified CM to accept other national numbering plans. Once this is done, the @ wildcard can be used for different numbering plans even within the same Unified CM cluster, depending on the value selected in the Numbering Plan field on the Route Pattern configuration page. For more information, please refer to the *Cisco Unified Communications Manager Dial Plan Deployment Guide*, available at

  https://www.cisco.com/en/US/products/sw/voicesw/ps5629/prod_maintenance_guides_list.html
- The @ wildcard can be practical in several small and medium deployments, but it can become harder to manage and troubleshoot in large deployments because adopting the @ wildcard forces the administrator to use route filters to block certain patterns (see Route Filters, page 14-27).

Route Filters

- Use route filters only with the @ route pattern to reduce the number of route patterns created by the @ wildcard. A route filter applied to a pattern not containing the @ wildcard has no effect on the resulting dial plan.
- The logical expression you enter with the route filter can be up to 1024 characters, excluding the NOT-SELECTED fields.
- As you increase the number of logical clauses in a route filter, the refresh time of the configuration page also increases and can become unacceptably long.
- When you configure call routing, be careful not to assign a single route filter to too many route patterns. A system core could result if you were to edit a route filter that has hundreds of associated route patterns. This is due to the extra system processing that is required to update call routing for all of the route patterns that use the route filter. Create duplicate route filters to ensure that this does not happen.
- For large-scale deployments, use explicit route patterns rather than the @ wildcard and route filters. This practice also facilitates management and troubleshooting because all patterns configured in Unified CM are easily visible from the Route Pattern configuration page.

International and Variable-Length Route Patterns

- International destinations are usually configured using the ! wildcard, which represents any quantity of digits. For example, in North America the route pattern 9.011! is typically configured for international calls. In most European countries, the same result is accomplished with the 0.00! route pattern.
- The ! wildcard is also used for deployments in countries where the dialed numbers can be of varying lengths. In such cases, Unified CM does not know when the dialing is complete and will wait for 15 seconds (by default) before sending the call. You can reduce this delay in any of the following ways:
  - Reduce the T302 timer (Service Parameter TimerT302_msec) to indicate end of dialing, but do not set it lower than 4 seconds to prevent premature transmission of the call before the user is finished dialing.
  - Configure a second route pattern followed by the # wildcard (for example, 9.011!# for North America or 0.00!# for Europe), and instruct the users to dial # to indicate end of dialing. This action is analogous to hitting the "send" button on a cell phone.

Overlap Sending and Overlap Receiving

In countries whose national numbering plan is not easily defined with static route patterns, you can configure Unified CM for overlap sending and overlap receiving.

Overlap sending means that Unified CM keeps collecting digits as they are dialed by the end users, and passes them on to the PSTN as they are dialed. To enable overlap sending, check the Allow Overlap Sending box on the Route Pattern Configuration page. The route pattern needs to include only the PSTN access code (for example, "9." in North America or "0." in many European countries).

Overlap receiving means that Unified CM receives the dialed digits one-by-one from a PRI PSTN gateway, and it then waits for completion of the dialed string before attempting to route the call to an internal destination. To enable overlap receiving, set the OverlapReceivingFlagForPRI service parameter to True.

Digit Manipulation in Route Patterns

- Digit manipulations configured on a route pattern affect the calling and called party number, no matter what route group the call eventually takes. Digit manipulations configured in the route list's view of its member route groups have a route-specific effect: only the transformations configured on the route group used to place the call will be performed.
- Digit manipulation in the route list's view of its member route group overrides any digit manipulation done in the route pattern.
- Transformation patterns configured on the device selected to route the call (or on that device's device pool) take precedence over calling and called party transformations configured in the route pattern and/or route list. If a transformation calling search space (CSS) is configured on the device selected to route the call (or on that device's device pool), then transformations configured in the route pattern or route list are considered only if no match is found using the respective transformation CSS. The input to the transformation CSS always is the untransformed number before applying route pattern or route list transformations.
- If you configure digit manipulation in the route pattern, the Call Detail Record (CDR) records the dialed number after the digit manipulation has occurred. If you configure digit manipulation only in the route group or on the device level, the CDR records the actual dialed number prior to the digit manipulation.
- Similarly, if you configure digit manipulation in the route pattern, the IP phone display of the calling party will show the manipulated number. If you configure digit manipulation only in the route group, the manipulations will be transparent to the end user.

Calling Line ID

- The calling line ID presentation can be enabled or disabled on the gateway and also can be manipulated in the route pattern, based on site requirements.
- If you select the option Use Calling Party's External Phone Number Mask, then the external call uses the calling line ID specified for the IP phone placing the call. If you do not select this option, then the mask placed in the Calling Party Transform Mask field is used to generate the calling party ID.

Call Classification

- Calls using this route pattern can be classified as on-net or off-net calls. This route pattern can be used to prevent toll fraud by prohibiting off-net to off-net call transfers or by tearing down a conference bridge when no on-net parties are present. (Both of these features are controlled by Service Parameters within Unified CM Administration.)
- When the "Allow device override" box is enabled, the calls are classified based on the call classification settings on the associated gateway or trunk.

Forced Authorization Codes (FAC)

- The Forced Authorization Codes (FAC) checkbox is used to restrict the outgoing calls when using a particular route pattern. If you enable FAC through route patterns, users must enter an authorization code to reach the intended recipient of the call.
- When a user dials a number that is routed through a FAC-enabled route pattern, the system plays a tone that prompts for the authorization code. To complete the call, the user authorization code must meet or exceed the level of authorization that is specified to route the dialed number.
- Only the authorization name appears in the call detail records (CDR); the authorization code does not appear in the CDR.
- The FAC feature is not available if the "Allow overlap sending" checkbox is enabled.

Client Matter Codes (CMC)

- The Client Matter Code (CMC) checkbox is used to track calls to certain numbers when using a particular route pattern. (For example, a company can use it to track calls to certain clients.)
- If you enable CMC for a route pattern, users must enter a code to reach the intended destination.
- When a user dials a number that is routed through a CMC-enabled route pattern, the system plays a tone that prompts for the code. The user must enter the correct code in order to complete the call.
- Client Matter Codes appear in the call detail records so that they can be used by the CDR analysis and reporting tool to generate reports for client billing and accounting.
- The CMC feature is not available if the "Allow overlap sending" checkbox is enabled.
- If both CMC and FAC are enabled, the user dials a number, enters the FAC when prompted to do so, and then enters the CMC at the next prompt.

SIP Route Pattern

SIP route patterns are configured in Unified CM to route or block calls to external entities based on the host portion (right-hand side) of SIP URIs. A SIP route pattern can point directly to a SIP trunk or point to a route list that then refers to one or more route groups and finally to SIP trunks. The use of the full SIP route pattern, route list, route group construct is highly recommended because it offers more flexibility.

SIP route patterns matching on the host piece of a SIP URI can match on a domain name or an IP address, both of which are possible as the right-hand side of a SIP URI. Wildcards can be used in domain name SIP route patterns to match on multiple domains (for example, *.cisco.com and ccm[1-4].uc.cisco.com). In IP address SIP route patterns, a subnet notation can be used (for example, 192.168.10.0/24).

Route Lists

A route list is a prioritized list of eligible paths (route groups) for an outbound call. A typical use of a route list is to specify two paths for a remote destination, where the first-choice path is across the IP WAN and the second-choice path is through a PSTN gateway.

Route lists have the following characteristics:

- Multiple route patterns may point to the same route list.
- A route list is a prioritized list of route groups that function as alternate paths to a given destination. For example, you can use a route list to provide least-cost routing, where the primary route group in the list offers a lower cost per call and the secondary route group is used only if the primary is unavailable due to an "all trunks busy" condition or insufficient IP WAN resources.

- Each route group in the route list can have its own digit manipulation. For example, if the route pattern is 9.@ and a user dials 9 1 408 555 4000, the IP WAN route group can strip off the 9 1 while the PSTN route group may strip off just the 9.
- Multiple route lists can contain the same route group. The digit manipulation for the route group is associated with the specific route list that points to the route group.
- If you are performing several digit manipulations in a route pattern or a route group, the order in which the transformations are performed can impact the resulting calling and called party numbers used for the call. Unified CM performs the following major types of digit manipulations in the order indicated:
  1. Discarding digits
  2. Called and calling party transformations as defined in the route pattern or for the route group
  3. Prefixing digits

Keep in mind that calling and called party transformations defined on the egress device (gateway or trunk) override calling and called party transformations defined in route patterns and route groups.

Route Groups

Route groups control and point to specific devices, which are typically gateways (MGCP, SIP, or H.323), H.323 or SIP trunks to a gatekeeper, remote Unified CM cluster, or Cisco Unified Border Element. Unified CM sends calls to the devices according to the distribution algorithm assigned. Unified CM supports top-down and circular algorithms.

Route Group Devices

The route group devices are the endpoints accessed by route groups, and they typically consist of gateways or trunks to a gatekeeper or to remote Unified CMs. You can configure the following types of devices in Unified CM:

- Media Gateway Control Protocol (MGCP) gateways
- SIP gateways
- H.323 gateways
- H.225 trunk, gatekeeper controlled — trunk to standard H.323 gateways, via a gatekeeper
- Intercluster trunk, not gatekeeper controlled — direct trunk to another Unified CM cluster
- Intercluster trunk, gatekeeper controlled — trunk to other Unified CM clusters and/or H.323 gateways, via a gatekeeper
- SIP trunk — trunk to another Unified CM cluster, a Cisco Unified Border Element, a Session Border Controller, or a SIP proxy

Note Both the H.225 and intercluster trunk (gatekeeper controlled) will automatically discover if the other endpoint is a standard H.323 gateway or a Unified CM and will select H.225 or Intercluster Trunk protocol accordingly.

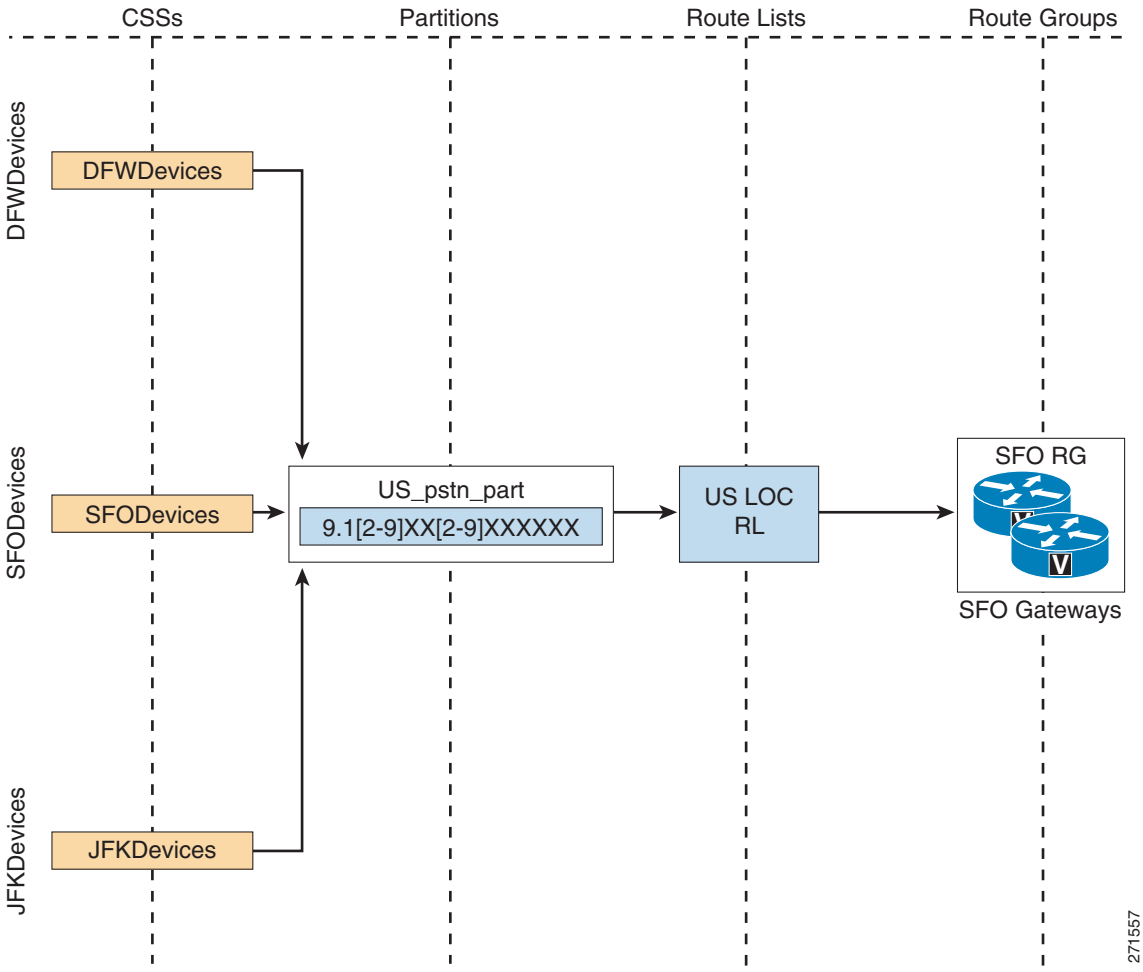
Local Route Group

Device pools can be associated with multiple local route groups. Route patterns using a local route group offer a unique characteristic: they allow for dynamic selection of the egress gateway, based on the device originating the call. By contrast, calls routed by route patterns using static route groups will route the call to the same gateway, no matter what device originated the call.

Example 14-2 Comparison of Local and Non-Local Route Groups

In Figure 14-15, a route pattern defined as 9.1[2-9]XX[2-9]XXXXXX points to a route list referencing a non-local route group containing San Francisco gateways. If this route pattern is placed in a partition contained in the calling search spaces of phones in Dallas, San Francisco, and New York, national calls from devices in those three cities will egress to the PSTN in San Francisco.

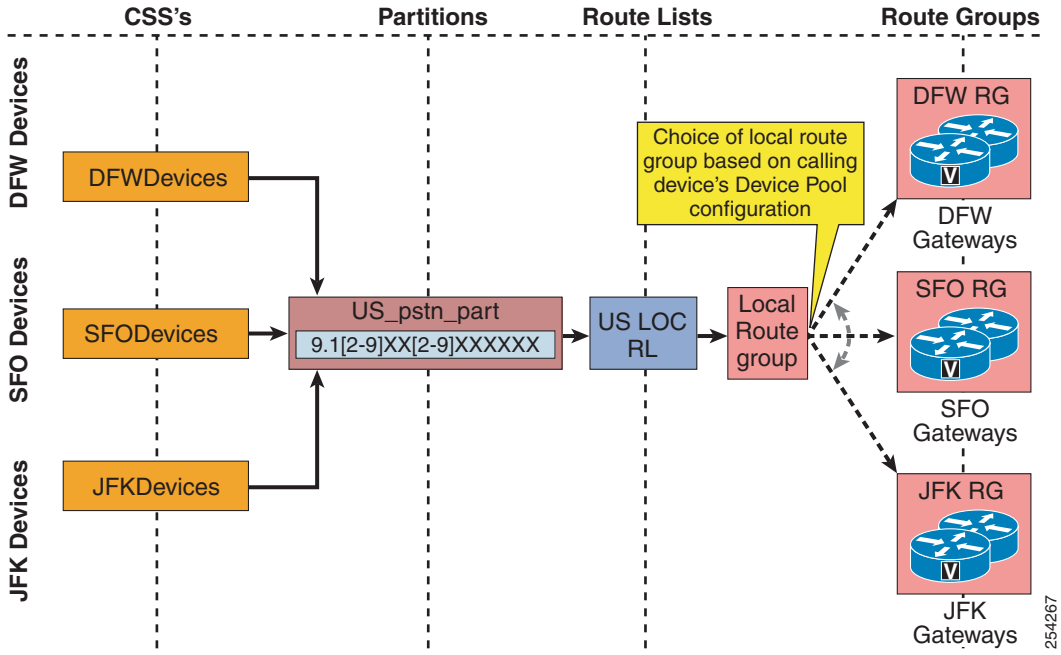
Figure 14-15 Non-Local Route Group Behavior



By contrast, if this same route pattern is modified to point to a route list containing the Standard Local Route Group as in Figure 14-16, then calls made from the Dallas site would egress to the PSTN through the Dallas gateway, calls made from the New York site would egress to the PSTN through the New York gateway, and calls made from the San Francisco site would egress to the PSTN through the San Francisco gateway.

The use of Local Route Group allows for egress gateway selection based on the calling device, which allows for site-independent route patterns that can be reused by calling search spaces for phones in all sites.

Figure 14-16 Local Route Group Behavior



The Device Mobility feature allows the device pool to be assigned to an endpoint based on the current subnet to which it has roamed. This permits assignment of the local route group to be based on the site where the phone is currently located.

Example 14-3 Device Mobility

A phone is moved from the San Francisco site to the New York site. Based on the phone's new IP address (part of the IP subnet associated with the New York site), a New York device pool is assigned to the phone. If the next call placed by the roaming phone matches a route pattern using a route list containing the Standard Local Route Group, it will be routed through the New York gateway.

If a local route group is used in forwarded call scenarios where, for example, phone A calls phone B and B is forwarded to a destination in the PSTN, then the route pattern in the call forward calling search space of phone B determines the class of service for calls forwarded by phone B, whereas by default the local route group associated with phone A's device pool is used to determine the egress gateways when hitting Standard Local Route group in the route list selected by the route pattern found using phone B's call forward calling search space. As a result, typically a gateway local to phone A is used for the forwarded call. This makes sure that the caller ID of the initial caller (phone A) can be sent to the PSTN

and that this caller ID will not be screened by the provider. There is a service parameter that allows administrators to configure the local route group selection policy for forwarded calls. The service parameter can be set to:

- **Calling Party's Local Route Group** — Backward compatible default. The local route group associated with the initial caller's device pool is selected (phone A in above example).
- **Original Called Party** — The local route group associated with the called phone's device pool is selected (phone B in above example).
- **Last Redirecting Party** — The local route group associated with the phone's device pool that is forwarding the call to the PSTN is selected (phone B in above example). These last two options differ only in cases where the call is forwarded through multiple hops before it finally gets forwarded out to the PSTN.

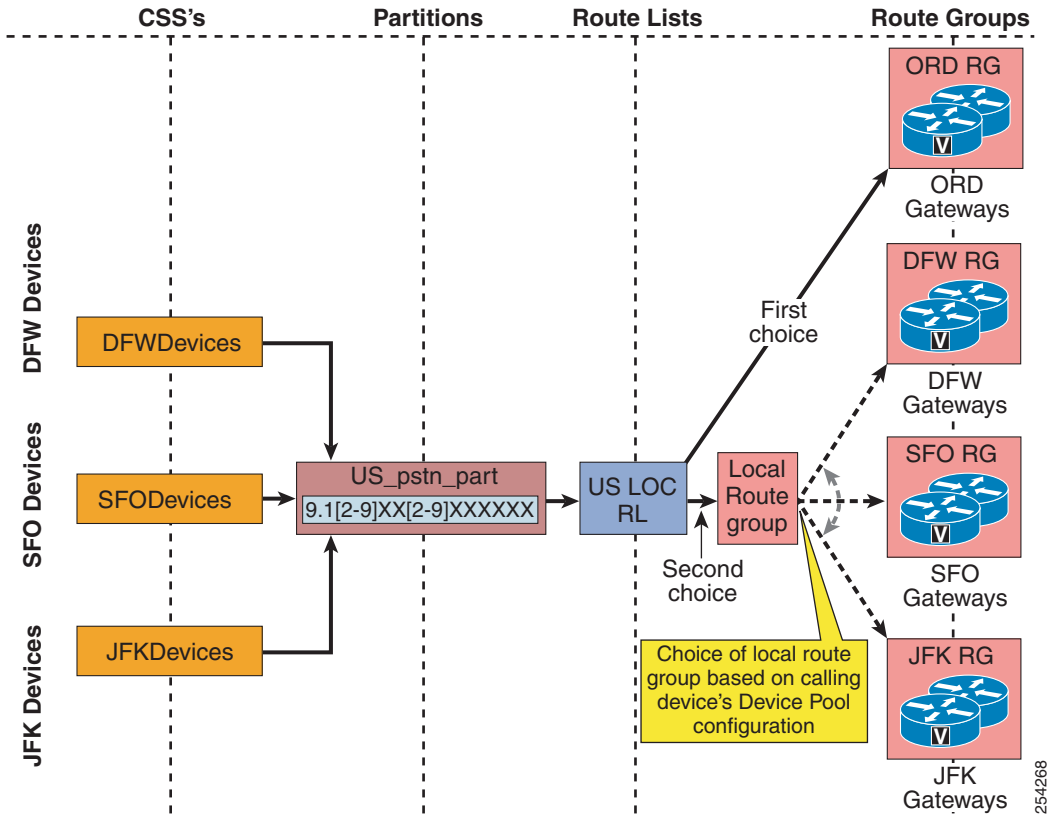
Centralized Gateway with Local Failover to the PSTN

Local route groups simplify the local failover to the PSTN for systems where a centralized gateway is configured. A single route list can be used to route PSTN calls for multiple sites while allowing local failover to the gateway at the site of origin.

Example 14-4 Centralized Gateways and Local Failover

A company negotiates a favorable PSTN interconnection rate for a group of trunks located in Chicago. If a route list includes a route group containing gateways in Chicago as its first entry and the Standard Local Route Group as the second choice, then any call it processes will first be sent to the preferred lower-cost gateways in Chicago. If a Chicago gateway is not available, if no ports are free, or if there is not enough bandwidth to allow the call between the calling phone and the Chicago gateway, then the next choice will be to attempt to route the call through the gateway co-located with the calling phone, as determined by the local route group in the calling phone's device pool configuration. (See Figure 14-17.)

Figure 14-17 Centralized Gateway with Local Failover to the PSTN



Multiple Local Route Groups

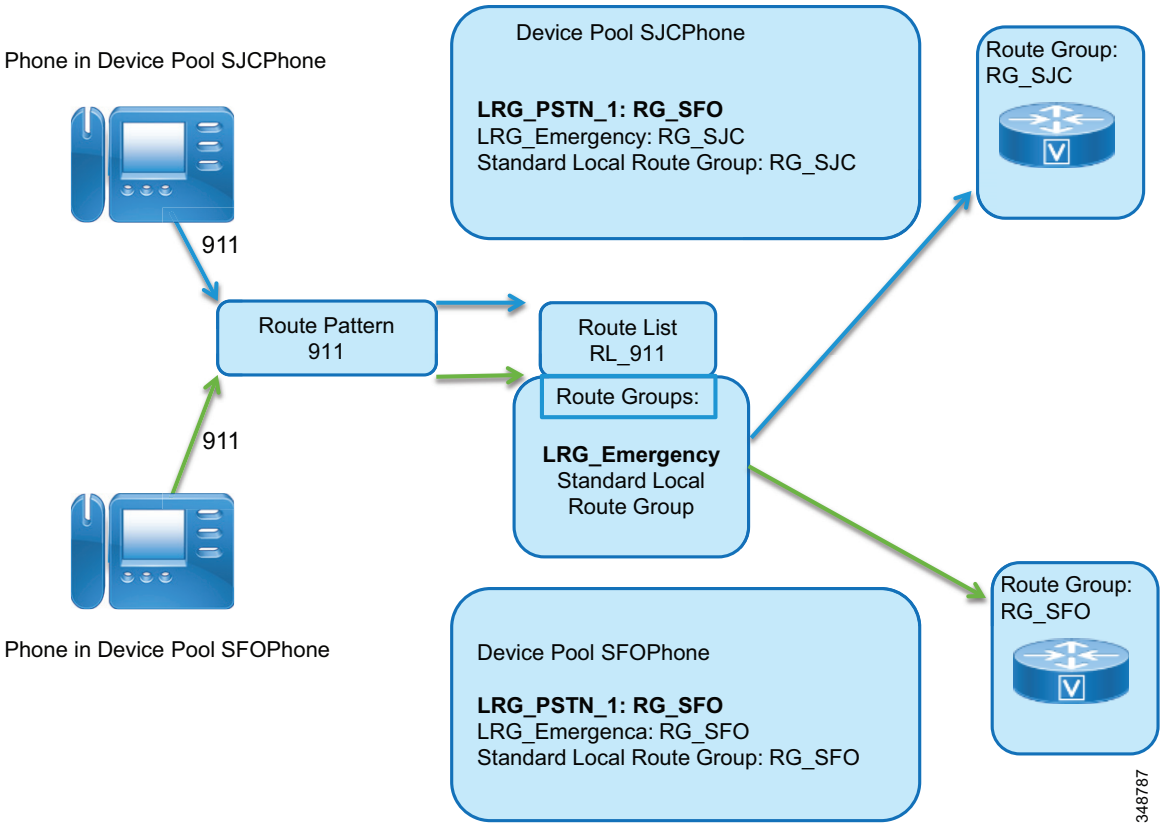
To support route lists with multiple route group elements specific to the calling device, multiple named local route groups can be configured in Unified CM. After names for all local route groups have been defined on the system level, a route group per named local route group can be configured on the device pool level. This, for example, allows to define different local route groups to be used for emergency calls, national PSTN destinations, and other destinations. Using multiple local route groups enables different gateways to be selected for different types of calls. For example, if small sites have small PSTN gateways that should be used only for emergency calls while PSTN calls of this small site should use the PSTN resources of a major hub, then we might want to use the following local route group configuration:

Sites	Local Route Groups	
	LRG_PSTN	LRG_Emergency
SJC (branch)	RG_SFO	RG_SJC
OAK (branch)	RG_SFO	RG_OAK
SFO (hub)	RG_SFO	RG_SFO
TPA (branch)	RG_MCO	RG_TPA
MIA (branch)	RG_MCO	RG_MIA
MCO (hub)	RG_MCO	RG_MCO

In this example the gateways in the major hubs (SFO and MCO) are used for PSTN calls by users in the hub sites and in the branch sites associated with the hub (SJC and OAK use SFO; TPA and MIA use MCO), while emergency calls always use local PSTN resources.

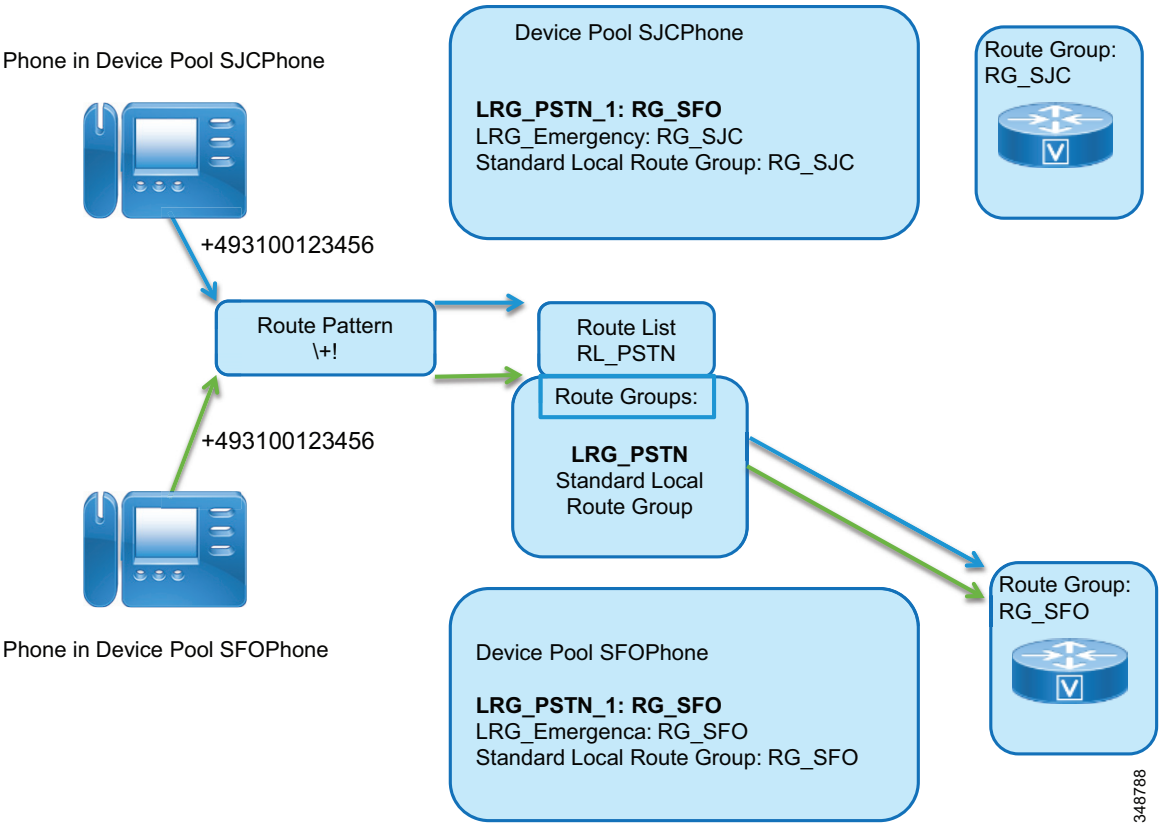
Figure 14-18 shows the call routing and local route group selection for an emergency call. Route list RL_911 used by the emergency route pattern would have LRG_Emergency as the first route group entry. The second entry in the route list refers to the Standard Local Route Group to make sure that the default PSTN resource defined on the device pool is selected as failover. Whenever an emergency call is placed and the route list entry LRG_Emergency is selected, Unified CM will dereference the placeholder LRG_Emergency and will instead use the route group configured for LRG_Emergency on the device pool of the calling device. The example shows how, for phones in sites SFO and SJC, local PSTN gateways are selected for emergency calls.

Figure 14-18 Emergency Call Routing with Multiple Local Route Groups



Using the same concept, a site-independent PSTN route pattern can be defined to point to a route list that uses LRG_PSTN. LRG_PSTN then is dereferenced to the route group defined on the device pool level for named local route group LRG_PSTN. Figure 14-19 shows how PSTN calls from sites SJC and SFO are routed to centralized PSTN gateways in site SFO, based on the device pool local route group settings.

Figure 14-19 PSTN Call Routing with Multiple Local Route Groups



Undefined local route groups are skipped during the egress routing device selection process. If a route list contains a local route group to which no route group has been assigned on the device pool of the calling device, then this entry in the route list is skipped and the next route group member of the route list is considered. When using route lists containing only local route groups, it is important to make sure that route groups are defined consistently on all device pools of all call originating devices to avoid dropping egress calls due to route list exhaustion without ever reaching a real route group.

Always using Standard Local Route Group as the last entry in all route lists and making sure that a route group for Standard Local Route Group is selected on all device pools, can be used as a safeguarding mechanism to avoid above route list exhaustion problem.

Pattern Urgency

Translation patterns, route patterns, and DNs can be configured as urgent patterns. The default value for pattern urgency is urgent for translation patterns and non-urgent for route patterns and DNs. Only the pattern urgency of route patterns, translation patterns, and DNs can be configured. All other patterns are always non-urgent.

Marking a pattern as urgent is often used to force immediate routing of certain calls as soon as a match is detected, without waiting for the T302 timer to expire. For example, in North America, if the patterns 9.911 and 9.[2-9]XXXXXX are configured and a user dials 9911, Unified CM has to wait for the T302 timer before routing the call because further digits may cause the 9.[2-9]XXXXXX pattern to match. If Urgent Priority is enabled for the 9.911 route pattern, Unified CM makes its routing decision as soon as the user has finished dialing 9911, without waiting for the T302 timer.

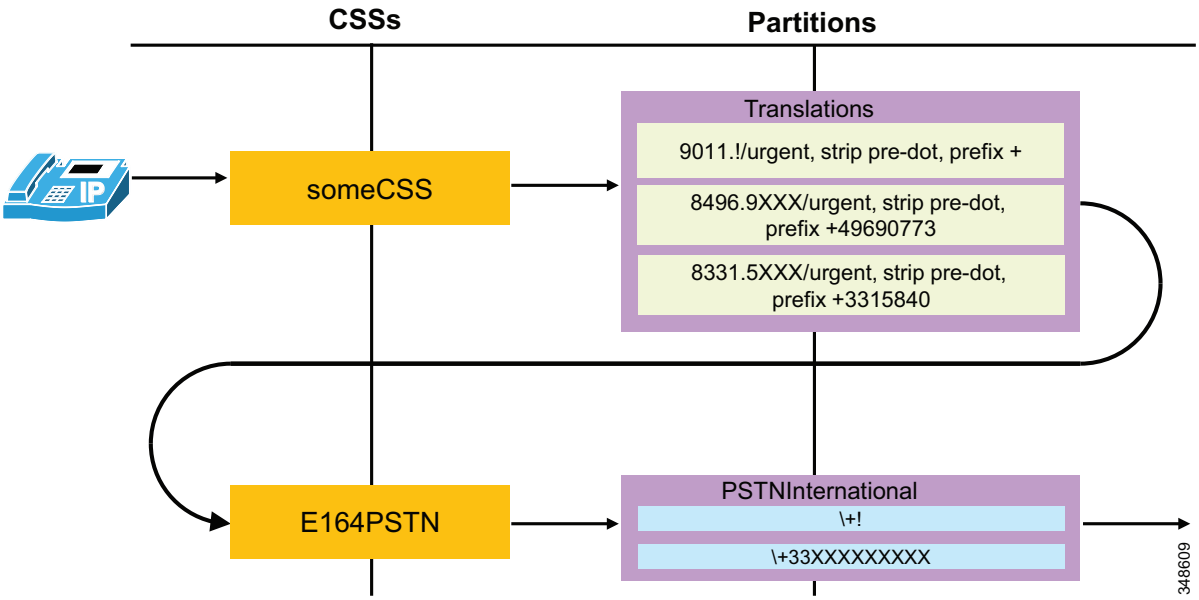
Making a pattern urgent forces the T302 timer to expire as soon as the configured pattern is the best match for the dialed number. This does not mean that the urgent pattern has a higher priority than other patterns; the closest-match logic described in the section on Call Routing in Unified CM, page 14-22, still applies.

For example, assume the route pattern 1XX is configured as urgent and the pattern 12! is configured as a non-urgent route pattern. If a user dials 123, Unified CM will not make its routing decision as soon as it receives the third digit because even though 1XX is an urgent pattern, it is not the best match (10 total patterns matched by 12! versus 100 patterns matched by 1XX). Unified CM will have to wait for inter-digit timeout before routing the call because the pattern 12! allows for more digits to be input by the user.

Consider another example where pattern 12[2-5] is marked as urgent and 12! is configured as a non-urgent pattern. If the user dials 123, the pattern 12[2-5] is the best match (four total patterns matched by 12[2-5] versus 10 patterns matched by 12!). Because the T302 timer is aborted and because the urgent-priority pattern is the best match, no further user input is expected. Unified CM routes the call using pattern 12[2-5].

A variable-length urgent translation pattern like 9011.! in Figure 14-20 will not force inter-digit timeout. As the dialed digits are received and analyzed digit-by-digit, as soon as an urgent translation pattern is the only (or best) match, the digit transformations defined on the translation pattern will be executed immediately and the secondary lookup as defined by the CSS on the translation pattern occurs.

Figure 14-20 Inter-Digit Timeout with Urgent Translations



Assuming the configuration in Figure 14-20, when the user dials 901133158405858 the call will be routed immediately after the last digit is dialed. The call will match translation pattern 9011.!, the dialed digits will be transformed to +3333158405858 (9011 discarded and + prefixed), which matches the fixed-length PSTN route pattern \+33XXXXXXXXX (nine-digit NSNs used in France).

On the other hand, if the user dials 9011496907739001, the user will experience inter-digit timeout. After matching 9011.! the resulting digits +496907739001 match route pattern \+!, and Unified CM needs to wait for further digits to make sure that the caller did not intent to continue to dial further digits. Further digits dialed would still match on the same route pattern.

The example in Figure 14-20 also shows how urgent translation patterns can be used to implement some abbreviated off-net dialing habit. Both translation patterns starting with 8 will accept exactly eight digits, transform the dialed digits to +E.164, and then execute the secondary lookup.

Dialing 83315858 will be routed immediately without inter-digit timeout. The dialed digits match fixed length translation pattern 8331.5XXX, and the translated called party number +33158405858 matches the fixed-length route pattern \+33XXXXXXXXX.

However, dialing 84969001 will not be routed immediately by default. The dialed digits are matched by the fixed-length translation pattern 8496.9XXX, and the translated called party number +496907739001 matches the variable-length PSTN route pattern \+!. This example shows that neither the pattern urgency nor the fixed-length characteristic of an intermediate translation pattern match is inherited by the secondary lookup defined by the CSS configured on the intermediate translation pattern (E164PSTN). Because the route pattern matched in the secondary lookup is a variable length pattern, Unified CM is forced to wait for inter-digit timeout. If the intermediate translation pattern is a fixed length translation pattern, waiting for further digits in the secondary lookup does not make much sense because any further digits will lead to a situation where the intermediate translation pattern will not be matched any more. Hence, for fixed length translation patterns it does make sense to change the inter-digit timeout handling for the secondary lookup. To achieve this, the option **Do Not Wait For Interdigit Timeout On Subsequent Hops** on the translation pattern has to be set. If this option is set, then after matching the translation pattern, Unified CM will not wait for any further digits and will just match the translated called party number against the patterns identified by the CSS defined on the intermediate translation pattern. As a general rule, **Do Not Wait For Interdigit Timeout On Subsequent Hops** should be enabled on all fixed length translation patterns.

Another typical use case for the **Do Not Wait For Interdigit Timeout On Subsequent Hops** option is the secondary lookup of dialing normalization translation patterns using a special key to terminate digit collection to avoid interdigit timeout. As an example, in a US dial plan a dialing normalization translation pattern matching international destinations with termination character # (such as 9011.!#) can match on variable length international dialing and allows users to terminate dialing by pressing #. This translation pattern's secondary lookup would typically match on a variable length route pattern such as \+[^1]! and this match in the secondary lookup would again force digit analysis to wait for further digits. Again the easiest way to avoid this timeout is to set the **Do Not Wait For Interdigit Timeout On Subsequent Hops** option on the dialing normalization translation pattern 9011.!#.

Calling and Called Party Transformation Patterns

A calling party transformation pattern allows the system to adapt the global form of the calling party's number into the local form required by off-cluster networks connected to the route group devices, such as gateways or trunks.

A called party transformation pattern allows the system to adapt the global form of the called party's number into the local form required by off-cluster networks connected to the route group devices.

Note Called party transformation patterns do not have any effect on phones. The called party transformation pattern CSS of the device pool does not impart any effects on the phones to which it is assigned.

Both pattern types consist of a numerical representation of the calling or called party number to be matched. The syntax used is the same as that of other patterns such as route patterns, translation patterns, directory numbers, and so forth. (See Figure 14-21.)

The transformation operators include discard digit instructions (for example, pre-dot), a calling party transformation mask, prefix digits, and control over the calling party presentation (either Default, Allowed, or Restricted). Calling party transformation patterns can be configured to use the calling party's external phone number mask as the calling party number.

Partitions and calling search spaces control which calling party transformation patterns are applied to which gateways or trunks. Gateways or trunks can use either their associated device pool's calling party transformation CSS or the device's own calling party transformation CSS, in reverse order of precedence. The same mechanism is used to control the applicability of called party transformation patterns.

Calling and called party transformation patterns configured on a Gateway Configuration page under **Call Routing Information - Outbound Calls** affect the calling or called party number sent to the gateway, as well as the calling or called party's numbering type and numbering plan. Calling and called party transformation patterns applied under **Incoming Calling Party Settings** apply to calls coming from the gateway.

Figure 14-21 Calling and Called Party Transformation Patterns

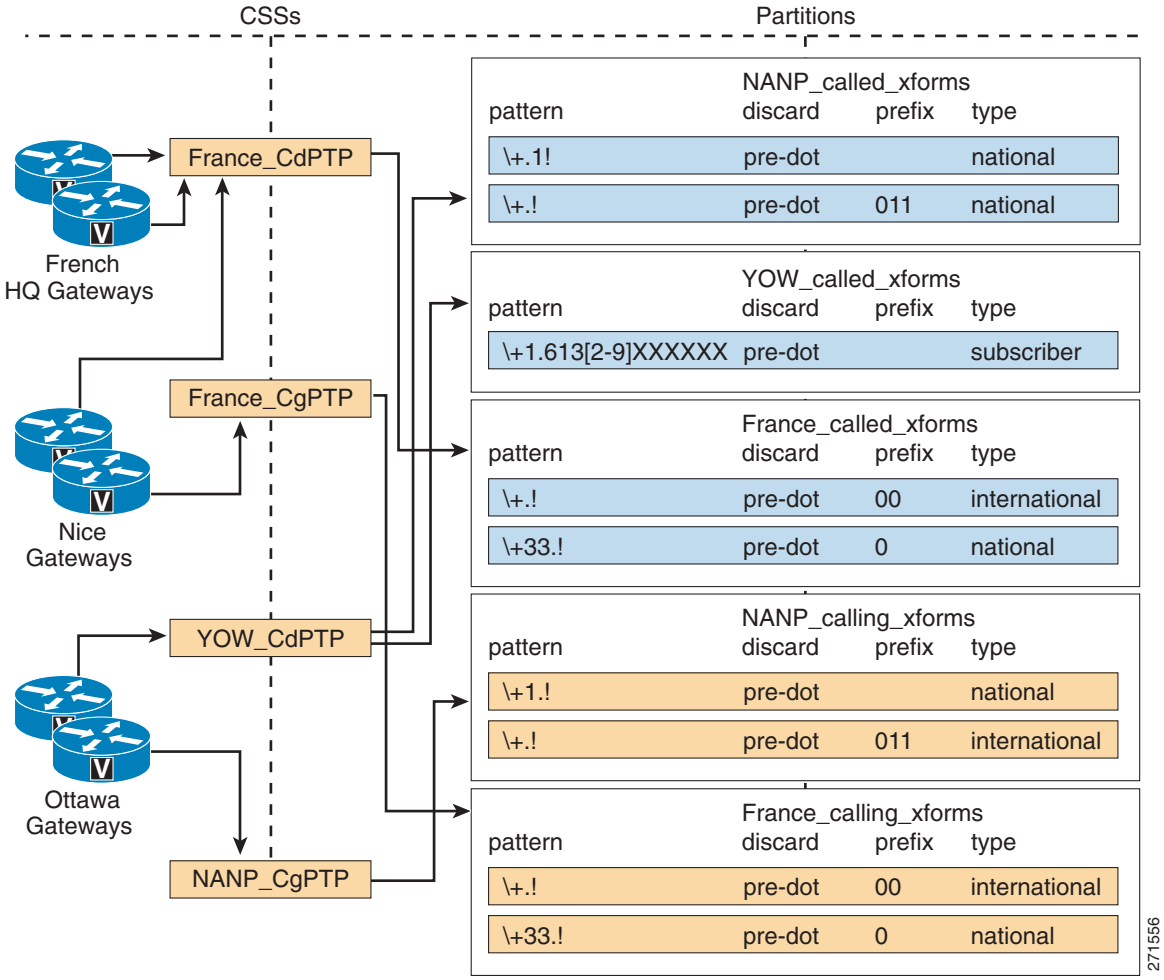


Figure 14-21 illustrates how calling and called party transformation patterns would be applied to different groups of gateways connected to the PSTN in different parts of the PSTN (France and NANP area).

Within the North American Numbering Plan (NANP), gateways located in Ottawa, Canada (airport code YOW) are assigned to the Calling Party Transformation CSS NANP_CgPTP, which contains partition NANP_calling_xforms. Any call with a calling party number beginning with +1 (that is, originating from within the NANP) would match both patterns configured within partition NANP_calling_xforms. Following the best-match logic, the first pattern will be chosen, and the calling party number will be stripped of the + sign and NANP country code 1. The remaining part of the calling party number will be used as the calling party number sent to the PSTN, with numbering type set to National.

For example, if a call from +1 613 555 1234 were sent out the YOW gateways, the calling party number would be transformed to 613 555 1234 with a numbering type set to National.

If a call from the same caller were to be sent to a French gateway, a different set of calling party transformation patterns would apply. For example, if a call from +1 613 555 1234 were sent out a gateway located in Nice, France (airport code NCE), the calling party transformation patterns contained in partition France_calling_xforms would be applied. In this case, the calling party number would be transformed to 001 613 555 1234 with the numbering type set to International.

Note Calling party number transformations may be overridden once the call is sent out the gateway. Many service providers will not permit calling party numbers outside a given range, as determined by local service agreements or regulations.

The same process applies to the called party number transformation patterns. For Ottawa gateways, the assigned called party transformation CSS is YOW_CdPTP, which contains partitions NANP_Called_xforms and YOW_Called_xforms. A call placed to a destination number within the Numbering Plan Area 613 would match all patterns contained in these two partitions. However, the best match process would select pattern \+1.613[2-9]XXXXXX.

For example, on a call placed to +1 613 555 9999 through the Ottawa gateways, the called party number would be transformed to 516 555 9999 with a numbering type set to Subscriber.

Incoming Calling Party Settings (per Gateway or Trunk)

Incoming calling party settings can be configured on individual gateways or trunks, at the device pool level, or at the service parameter level, in order of precedence. For each numbering type (Subscriber, National, International, or Unknown), Unified CM allows for the appropriate prefix digits to be configured. Using the service parameter settings is not recommended because the device pool and gateway or trunk settings also allow for definition of strip digit instructions and flexible transformations based on calling party transformation patterns. Digits can be stripped from and prefixed to the string provided as the incoming party number. The digit stripping operation is performed first on the incoming calling party number, and then the prefix digits are added to the resulting string. For example, if the number of digits to be stripped is set to 1 and the prefix digits are set to +33, and the incoming calling party number is 01 58 40 58 58, then the resulting string will be +33 1 58 40 58 58.

Each numbering type can be configured with a calling search space used to apply calling party transformation patterns to the calls. The calling search space should contain partitions containing calling party transformation patterns exclusively. This allows the modifications applied to the calling party number to be based on the structure of the calling party number rather than strictly on its numbering type. The calling party transformation patterns use regular expressions to match the calling party number. The best-match process is used to choose between multiple matches, and the selected pattern's calling party transformations are applied to the call.

Incoming Called Party Settings (per Gateway or Trunk)

Equivalent to the incoming calling party settings described in the previous section, incoming called party transformations can also be configured. These incoming called party transformations enable normalization of incoming called party information prior to actually routing the call.

Each numbering type can be configured with a calling search space used to apply called party transformation patterns to the calls. The calling search space should contain partitions containing called party transformation patterns exclusively. This allows the modifications applied to the called party number to be based on the structure of the called party number rather than strictly on its numbering type. The called party transformation patterns use regular expressions to match the called party number. The best-match process is used to choose between multiple matches, and the selected pattern's called party transformations are applied to the call.

Calling Privileges in Unified CM

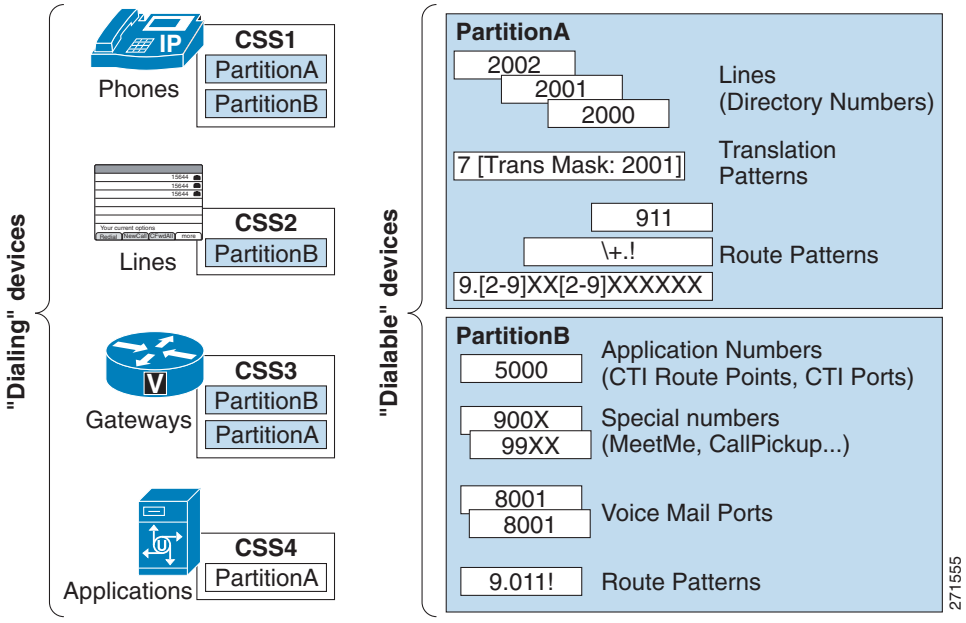
Dialing privileges are configured in order to control which types of calls are allowed (or prevented) for a particular endpoint (such as phones, gateways, or CTI applications). All calls handled by Unified CM are subjected to the dialing privileges implemented through the configuration of the following elements:

- Partitions, page 14-42
- Calling Search Spaces, page 14-43

A *partition* is a group of directory numbers (DNs) or directory URIs with similar accessibility, and a *calling search space* defines which partitions are accessible to a particular device. A device can call only those DNs and directory URIs located in the partitions that are part of its calling search space.

As illustrated in Figure 14-22, items that can be placed in partitions all have a dialable pattern, and they include phone lines, route patterns, translation patterns, CTI route group lines, CTI port lines, voicemail ports, and Meet-Me conference numbers. Conversely, items that have a calling search space are all devices capable of dialing a call, such as phones, phone lines, gateways, and applications (via their CTI route groups or voicemail ports).

Figure 14-22 Partitions and Calling Search Spaces

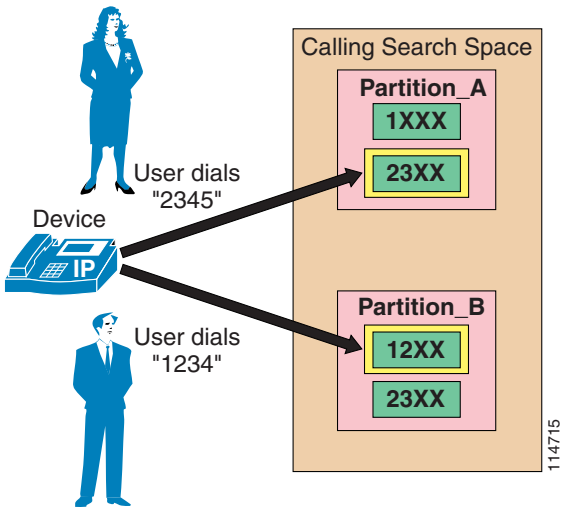


Partitions

The dial plan entries that you may place in a partition include IP phone directory numbers, directory URIs, translation patterns, route patterns, CTI route points, and voicemail ports. As described in the section on Call Routing in Unified CM, page 14-22, if two or more numeric dial plan entries (directory numbers, route patterns, or so forth) overlap, Unified CM selects the entry with the closest match (most specific match) to the dialed number. In cases where two dial plan entries match the dialed pattern equally, Unified CM selects the dial plan entry that appears first in the calling search space of the device making the call. Directory URIs always have to match completely; there is no concept of partial matches for directory URIs.

For example, consider Figure 14-23, where route patterns 1XXX and 23XX are part of Partition_A and route patterns 12XX and 23XX are part of Partition_B. The calling search space of the calling device lists the partitions in the order Partition_A:Partition_B. If the user of this device dials 2345, Unified CM selects route pattern 23XX in Partition_A as the matching entry because it appears first in the calling device's calling search space. However, if the user dials 1234, Unified CM selects route pattern 12XX in Partition_B as the matching entry because it is a closer match than 1XXX in Partition_A. Remember that the partition order in a calling search space is used exclusively as a tie-breaker in case of equal matches based on the closest-match logic.

Figure 14-23 Impact of Partition Order on the Matching Logic



Note When multiple equal-precision matches occur in the same partition, Unified CM selects the entry that is listed first in its local dial plan database. Since you cannot configure the order in which the dial plan database lists dial plan entries, Cisco strongly recommends that you avoid any possibility of equal-precision matches coexisting within the same partition because the resulting dial plan logic is not predictable in such cases.

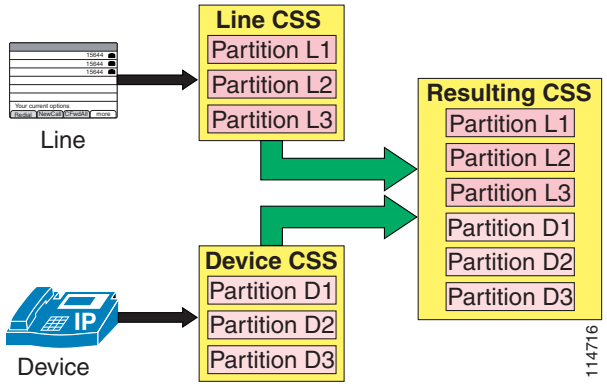
Partitions can be activated or deactivated based on the time and date. You can activate or deactivate partitions by first configuring time periods and schedules within Unified CM Administration and then assigning a specific time schedule to each partition. Outside of the times and days specified by the schedule, the partition is inactive, and all patterns contained within it are ignored by the Unified CM call routing engine. For more information on this feature, see Time-of-Day Routing, page 14-91.

Calling Search Spaces

A calling search space defines which partitions are accessible to a particular device. Devices that are assigned a certain calling search space can access only the partitions listed in that calling search space. Attempts to dial a DN or directory URI in a partition outside that calling search space will fail, and the caller will hear a busy signal.

If you configure a calling search space both on an IP phone line and on the device (phone) itself, Unified CM concatenates the two calling search spaces and places the line's calling search space in front of the device's calling search space, as shown in Figure 14-24.

Figure 14-24 Concatenation of Line and Device Calling Search Spaces for IP Phones



Note When device mobility is not used, the device calling search space is static and remains the same even as the device is moved to different parts of the network. When device mobility is enabled, the device calling search space can be determined dynamically based on where in the network the phone is physically located, as determined by the phone's IP address. See Device Mobility, page 14-83, for more details.

If the same pattern appears in two partitions, one contained in the line's calling search space and one contained in the device's calling search space, then according to the rules described in the section on Partitions, page 14-42, Unified CM selects the route pattern listed first in the concatenated list of partitions (in this case, the pattern associated with the line's calling search space).

The maximum length of any calling search space is 1024 characters, including separator characters between each partition name. (For example, the string “partition_1:partition_2:partition_3” contains 35 characters.) Thus, the maximum number of partitions in a calling search space varies, depending on the length of the partition names. Therefore, when you are creating partitions and calling search spaces, keep the names of partitions short relative to the number of partitions that you plan to include in a calling search space. For more details on configuring calling search spaces, refer to the *Cisco Unified Communications Manager Administration Guide*, available online at

https://www.cisco.com

Before you configure any partitions or calling search spaces, all DNs reside in a special partition named <None>, and all devices are assigned a calling search space also named <None>. When you create custom partitions and calling search spaces, any calling search space you create also contains the <None> partition, while the <None> calling search space contains *only* the <None> partition.

Note Any dial plan entry left in the <None> partition is implicitly reachable by *any* device making a call. Therefore, to avoid unexpected results, Cisco strongly recommends that you do not leave dial plan entries in the <None> partition.

Note Cisco strongly recommends that you do not leave any calling search space defined as <None>. Doing so can introduce dial plan behavior that is difficult to predict.

Special Considerations for Transformation Patterns

Calling and called transformation patterns are also placed in partitions, and those partitions are included in calling search spaces (CSSs) but not in order to control calling privileges. The partitioning of transformation patterns serves to choose which transformations are applied to which gateways, trunks, or phones. Partitions contained in calling party transformation pattern CSSs should contain only calling party transformation patterns. Likewise, partitions contained in called party transformation pattern CSSs should contain only called party transformation patterns.

Call-Forward Calling Search Spaces

Note Call Forward All actions are different than any other call-forward action in that the destination number is entered by each individual user when the feature is activated from a phone.

The system allows you to decide how call-forward calling search spaces take effect. There are three possible options, as selected by the Calling Search Space Activation policy:

- Use System Default

  If you configure the Calling Search Space Activation Policy to Use System Default, then the CFA CSS Activation Policy cluster-wide service parameter determines which Forward All Calling Search Space will be used. The CFA CSS Activation Policy service parameter can be set to With Configured CSS or to With Activating Device/Line CSS (see below). By default, the CFA CSS Activation Policy service parameter is set to With Configured CSS.

- With Configured CSS

  If you select the With Configured CSS option, the Forward All Calling Search Space and Secondary Calling Search Space for Forward All explicitly configured in the Directory Number Configuration window control the forward-all activation and call forwarding. If the Forward All Calling Search Space is set to None, no CSS gets configured for Forward All. A forward-all activation attempt to any directory number with a partition will fail. No change in the Forward All Calling Search Space and Secondary Calling Search Space for Forward All occurs during the forward-all activation.

- With Activating Device/Line CSS

  If you prefer to use the combination of the Directory Number Calling Search Space and Device Calling Search Space without explicitly configuring a Forward All Calling Search Space, select With Activating Device/Line CSS for the Calling Search Space Activation Policy. With this option, when Forward All is activated from the phone, the Forward All Calling Search Space and Secondary Calling Search Space for Forward All are automatically populated with the Directory Number Calling Search Space and Device Calling Search Space for the activating device. When you set the Forward All Destination from Unified CM Administration, the Forward All Calling Search Space

and Secondary Calling Search Space are not automatically populated and have to be configured explicitly. The two calling search spaces are concatenated, and the resulting calling search space is used to validate the number entered as a Call Forward All destination.

With this configuration (Calling Search Space Activation Policy set to With Activating Device/Line), if the Forward All Calling Search Space is set to None when forward-all is activated through the phone, the combination of Directory Number Calling Search Space and activating Device Calling Search Space is used to verify the forward-all attempt.

On Type-A IP phones running SIP, if Call Forward All is invoked from the phone itself, the device's Rerouting Calling Search Space is used for forwarded calls. If Forward All actions are invoked from the Unified CM User page or the Unified CM Administrative page, then any Forward All action initiated from the phone is irrelevant.

For example, assume an Type-A IP phone running SIP is configured with Forward All to extension 3000 from the Unified CM User page. At the same time, the phone itself is configured to Forward All to extension 2000. All calls made to that phone will be forwarded to extension 3000.

Note On Type-A IP phones running SIP, invoking Forward All from the Unified CM User or Administrative pages will not be reflected on the phone. The phone does not display any visual confirmation that calls are forwarded.

When Forward All is initiated from an IP phone running SCCP or from an Type-B IP phone running SIP, user input is simultaneously compared to the patterns allowed in the configured Forward All calling search space(s). If an invalid destination pattern is configured, the user will be presented with reorder tone. When Forward All is invoked from an Type-A IP phone running SIP, Forward All user input is stored locally on the phone and is not verified against any calling search space in Unified CM. If user input corresponds to an invalid destination, no notification is offered to the user. Calls made to that phone will be presented with reorder tone as the phone tries to initiate a SIP re-route action to an invalid destination number.

Other Call Forward Types

The calling search spaces configured for the various other types of call forward (Forward Busy, Forward No Answer, Forward No Coverage, forward on CTI failure, and Forward Unregistered) are standalone values not concatenated with any other calling search space.

Call Forward settings (except Forward All) can be configured separately for internal or external call types. For example, a user might want to have their phone Call Forward No Answer to voicemail for external callers but forward to a cell phone number if the caller is a co-worker calling from another IP phone on the network. This is possible by using different configurations for the Internal and External Call Forward settings.

When the Forward All calling search space is left as <None>, the results are difficult to predict and depend on the Unified CM release. Therefore, Cisco recommends the following best practices when configuring call-forward calling search spaces:

- Always provision the call-forward calling search spaces with a value other than <None>. This practice avoids confusion and facilitates troubleshooting because it enables the network administrator to know exactly which calling search space is being used for forwarded calls.
- Configure the Call Forward Busy and Call Forward No Answer calling search spaces with values that allow them to reach the DNs for the voicemail pilot and voicemail ports but not external PSTN numbers.

- Configure both the Call Forward All calling search space and the Secondary Calling Search Space for Forward All, according to your company's policy. Many companies choose to restrict forwarded calls to internal numbers only, to prevent users from forwarding their IP phone lines to a long-distance number and dialing their local IP phone number from the PSTN to bypass long-distance toll charges on personal calls.

The Call Forward Unregistered (CFUR) feature is a way to reroute calls placed to a temporarily unregistered destination phone. The configuration of CFUR consists of two main elements:

- Destination selection

  When the DN is unregistered, calls can be rerouted to either of the following destinations:

  - Voicemail

    Calls can be sent to voicemail by selecting the voicemail checkbox and configuring the CFUR calling search space to contain the partition of the voicemail pilot number.

  - A directory number used to reach the phone through the PSTN

    This approach is preferred when a phone is located within a site whose WAN link is down. If the site is equipped with Survivable Remote Site Telephony (SRST), the phone (and its co-located PSTN gateway) will re-register with the co-located SRST router. The phone is then able to receive calls placed to its PSTN DID number.

    In this case, the appropriate CFUR destination is the corresponding PSTN DID number of the original destination DN. Configure this PSTN DID in the destination field, preferably in E.164 format, including the + sign (for example, +1 415 555 1234). This allows the CFUR destination to be processed by the calling phone's local route group, whether or not it uses the same off-net access code and PSTN prefixes as the unregistered phone.

- Calling search space

  Unified CM attempts to route the call to the configured destination number by using the called DN's CFUR calling search space. The CFUR calling search space is configured on the target phone and is used by all devices calling the unregistered phone. This means that all calling devices will use the same combination of route pattern, route list, and route group to place the call. Cisco recommends that you configure the CFUR calling search space to route calls to the CFUR destination using patterns pointing to route lists referencing the Standard Local Route Group. This will ensure that the egress gateway to the PSTN is chosen based on the calling device.

The Call Forward Unregistered functionality can result in telephony routing loops if a phone is unregistered while the gateway associated with the phone's DID number is still under control of Unified CM, as is the case if a phone is simply disconnected from the network. In such a case, the initial call to the phone would prompt the system to attempt a first CFUR call to the phone's DID through the PSTN. The resulting incoming PSTN call would in turn trigger another CFUR attempt to reach the same phone's DN, triggering yet another CFUR call from the central PSTN gateway through the PSTN. This cycle could repeat itself until system resources are exhausted.

The service parameter **MaximumForwardUnRegisteredHopsToDn** controls the maximum number of CFUR calls that are allowed for a DN at the same time. The default value of 0 means the counter is disabled. If any DNs are configured to reroute CFUR calls through the PSTN, loop prevention is required. Configuring this service parameter to a value of 1 would stop CFUR attempts as soon as a single call is placed through the CFUR mechanism. This setting would also allow only one call to be forwarded to voicemail, if CFUR is so configured. Configuring this service parameter to a value of 2 would allow for up to two simultaneous callers to reach the voicemail of a DN whose CFUR setting is configured for voicemail, while also limiting potential loops to two for DNs whose CFUR configuration sends calls through the PSTN.

Note Extension Mobility DNs should not be configured to send Call Forward Unregistered calls to the PSTN DID associated with the DN. The DNs of Extension Mobility profiles in the logged-out state are deemed to be unregistered, therefore any calls to the PSTN DID number of a logged-out DN would trigger a routing loop. To ensure that calls made to Extension Mobility DNs in the logged-out state are sent to voicemail, ensure that their corresponding Call Forward Unregistered parameters are configured to send calls to voicemail.

Global Dial Plan Replication

With Global Dial Plan Replication (GDPR), independent Unified CM clusters can share dial plan elements such as URIs, +E.164 numbers, enterprise numbers, +E.164 patterns, enterprise patterns, and PSTN failover numbers using the Intercluster Lookup Service (ILS). All local dial plan information advertised by a Unified CM cluster is advertised as part of a single GDPR catalog. Reachability for advertised dial plan elements is achieved by advertising a location attribute (SIP route string) together with each GDPR catalog.

Enterprise-specific numbers and patterns represent a global enterprise-specific dialing habit that allows abbreviated on-net inter-site dialing. Enterprise-specific numbers and patterns to be exchanged through GDPR have to be globally significant. +E.164 numbers and patterns based on the characteristics of an E.164 numbering scheme are globally significant by definition.

This location attribute in multi-cluster environments is used to direct calls for any destination learned via GDPR to the correct cluster. For directory URIs this can be used when the host portion of the directory URI cannot be used to deterministically route the SIP request. This, for example, is the case when a flat URI scheme such as <user>@example.com is used. The host portion, example.com, does not uniquely identify the remote Unified CM cluster that hosts a given URI, but an appropriately chosen SIP route string does.

For every DN in Unified CM a +E.164 alternate number and an enterprise alternate number can be defined based on masks to be applied to the configured DN. These alternate numbers can optionally be added into local partitions. Each alternate number can be configured individually to be advertised to remote clusters using GDPR.

For every DN in Unified CM, up to five URIs can be defined as aliases. Each individual URI can be configured to be advertised to remote clusters using GDPR.

For every DN in Unified CM, the enterprise or +E.164 alternate number can be selected to be advertised as the PSTN failover number. On remote clusters this PSTN failover number is used for PSTN failover for calls to the +E.164 alternate number, enterprise alternate number, or URIs. PSTN failover is triggered if a call to any GDPR learned destination fails with cause codes other than **unallocated number**, **user busy**, **normal call clearing**, **destination out of order**, or **service not available**. The PSTN failover number is also used for Automated Alternate Routing (AAR) in case of call admission control failure. For calls to the PSTN failover number, the AAR CSS of the calling device is used on the remote cluster.

In addition to DN related information (directory URIs, enterprise alternate numbers, +E.164 alternate numbers, and PSTN failover numbers), GDPR also allows advertising of enterprise patterns and +E.164 patterns. Patterns are not associated with DNs and can be defined using wildcards (fixed length and variable length). The PSTN failover number for enterprise and +E.164 patterns is defined based on strip and prefix instructions.

GDPR not only allows advertising of local routing information but also supports imported GDPR catalogs that can contain URIs, enterprise patterns, and +E.164 patterns. For each imported GDPR catalog, a unique locations attribute (SIP route string) is advertised. This allows clusters to inject routing information for non-local destinations.

On the receiving side, all directory URIs learned through GDPR are put into a single local repository to be consulted when routing a non-numeric URI does not find a local URI match. All learned URIs are treated as equivalent from the class-of-service perspective.

In contrast to this, numeric patterns and numbers learned through GDPR are put into local partitions based on the type of information. Four separate partitions can be configured for +E.164 alternate numbers, enterprise alternate numbers, +E.164 patterns, and enterprise patterns. The default partitions for these different types of learned information are **Global Learned E164 Numbers**, **Global Learned E164 Patterns**, **Global Learned Enterprise Numbers**, and **Global Learned Enterprise Patterns**. To avoid unnecessary inter-digit timeout when dialing remote destinations learned through GDPR, the pattern urgency for learned destinations can be configured per class.

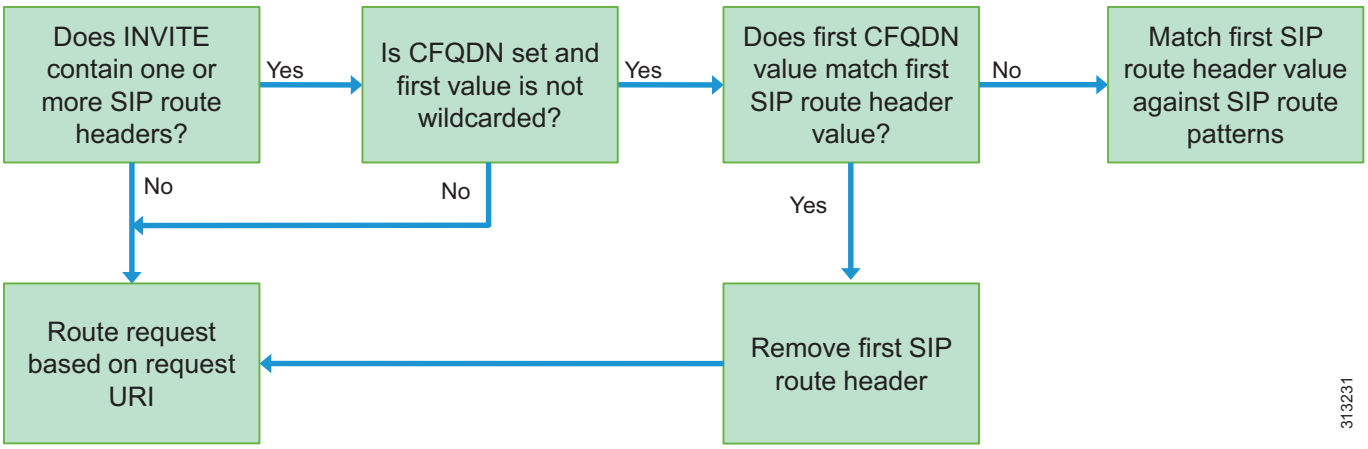
Cisco strongly recommends configuring +E.164 numbers and fixed length +E.164 patterns to be inserted into local digit analysis as **urgent**.

For details of how calls to directory URIs and numeric destinations learned through GDPR are routed, see the section on Routing of SIP Requests in Unified CM, page 14-48.

Routing of SIP Requests in Unified CM

Routing of SIP requests received from SIP trunks or SIP endpoints follows certain rules to make sure that both local and intercluster routing requirements are met. Figure 14-25 shows how Unified CM treats SIP route headers if any of them are present in a SIP request.

Figure 14-25 SIP Route Header-Based Routing



Before analyzing the request URI of a SIP request, Unified CM first checks for presence of a SIP route header (for example, Route: <sip:ucm.example.com;lr>). If no SIP route header is present, then Unified CM routes the SIP request based on the SIP request URI.

If one or more SIP route headers are present, if the Cluster Fully Qualified Domain Name (CFQDN) enterprise parameter is set, and if the first value in that parameter is not a wildcard, then Unified CM considers the first route header value in the first SIP route header for routing. Unified CM checks whether the host specification in the SIP route header value (ucm.example.com in the above example) matches the first entry in the Cluster Fully Qualified Domain Name enterprise parameter. If that is the case, then the topmost SIP route header is removed and Unified CM routes the request based on the request URI.

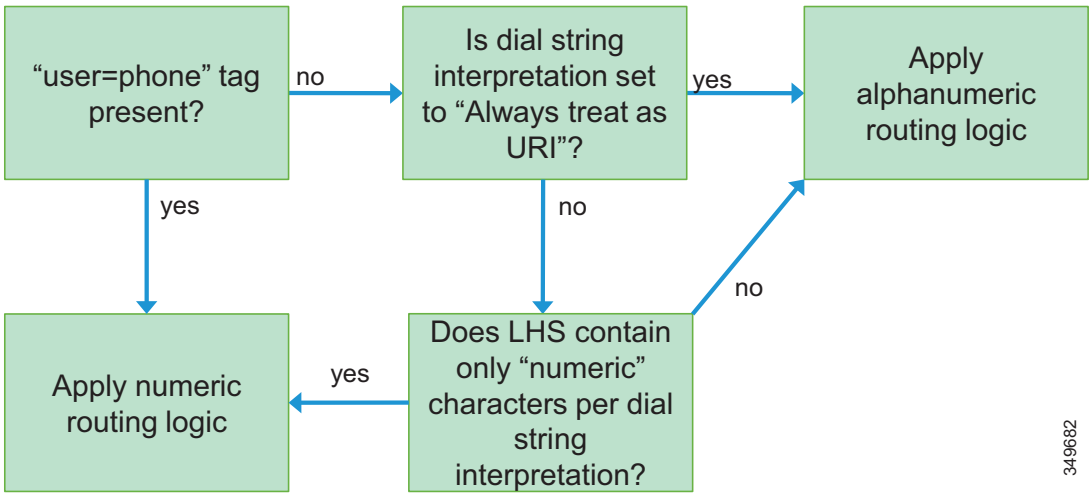
If a SIP route header is present and the host specification in the first SIP route header value does not match the first entry in the Cluster Fully Qualified Domain Name enterprise parameter, then Unified CM routes the request by matching the SIP route header value against the configured SIP route patterns. This routing behavior makes sure that Unified CM SME can be used as the transit routing entity between Cisco Expressway-C and other enterprise Unified CM clusters in Cisco Spark Hybrid Call Service deployments where for calls initiated on Cisco Spark applications by users configured for Call Service Connect a call leg is forked from the Cisco Collaboration Cloud to the calling user's Unified CM cluster to be anchored on the calling user's Cisco Spark Remote Device. The SIP request of this call leg carries the dialed destination in the request URI and the Cluster Fully Qualified Domain Name of the calling user's Unified CM cluster in a SIP route header added to the request by the Cisco Collaboration Cloud.

Unified CM applies different routing logic for numeric and alphanumeric SIP request URIs.

Numeric URI Versus Directory URI

Figure 14-26 shows the decision tree used by Unified CM to classify whether an incoming SIP request URI should be treated as an alphanumeric or a numeric URI.

Figure 14-26 Numeric vs. Alphanumeric URI Classification



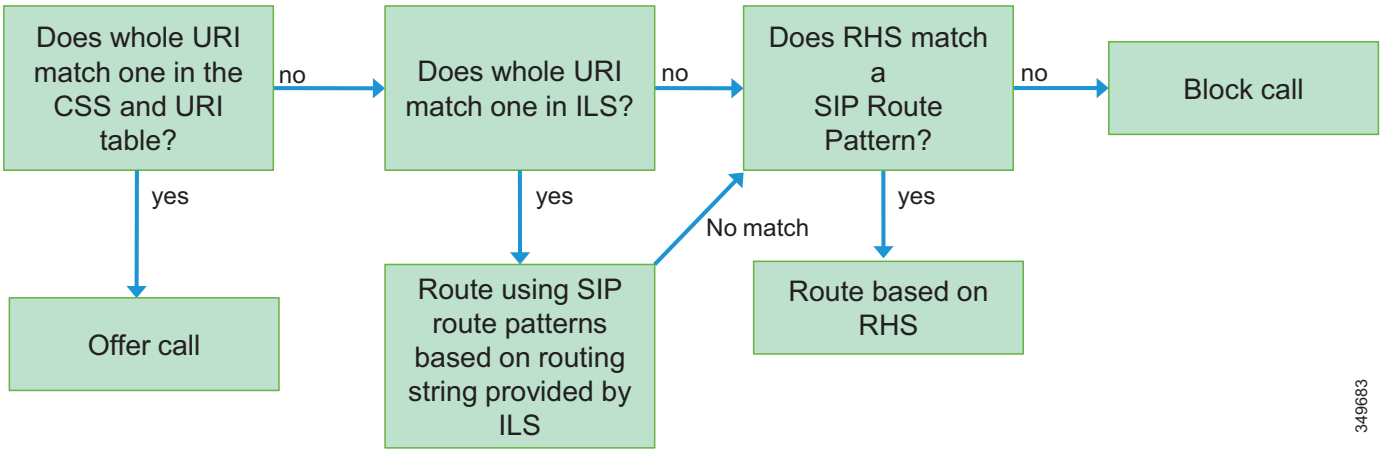
If the SIP request carries a user=phone tag, the SIP URI will always be interpreted as a numeric SIP URI. If no user=phone is present, the decision is based on the dial string interpretation setting in the calling device's (endpoint or trunk) SIP profile. This setting either defines a set of characters that Unified CM will accept as part of numeric SIP URIs (0-9, *, #, +, and optionally A-D) or it enforces the interpretation as a directory URI.

The routing logic applied to numeric and alphanumeric URIs is described in the following sections.

Routing Alphanumeric Directory URIs

Figure 14-27 shows a flowchart of the routing logic applied to alphanumeric URIs by Unified CM.

Figure 14-27 Call Routing Logic for SIP Request



The first step is to try to route the SIP request based on the calling search space of the calling device. Unified CM searches for a full match of the SIP URI against all directory URIs configured in the partitions addressed by the calling device's calling search space. If a match is found, the call is extended to the directory number associated with the matched local directory URI.

In case no matching local directory URI is found, Unified CM tries to locate the SIP URI in imported GDPR catalogs or GDPR catalogs learned from remote systems, again by searching for a full match. In case of a match, the SIP request is routed by matching the SIP route string (location identifier) associated with the GDPR catalog, as part of which the found directory URI was learned, against configured SIP route patterns addressed by the calling device's calling search space. (See Figure 14-28.)

In case the SIP URI does not match a local directory URI and also does not match any directory URI in any GDPR catalog, Unified CM then routes the SIP request based only on matching the right-hand side of the SIP URI against configured SIP route patterns. This routing of last resort can be used to create a default route for all SIP URIs not known locally or on any other call control participating in GDPR. A typical example for this is a SIP route to a Cisco Expressway business-to-business (B2B) building block.

Figure 14-28 Example for Routing a Directory URI

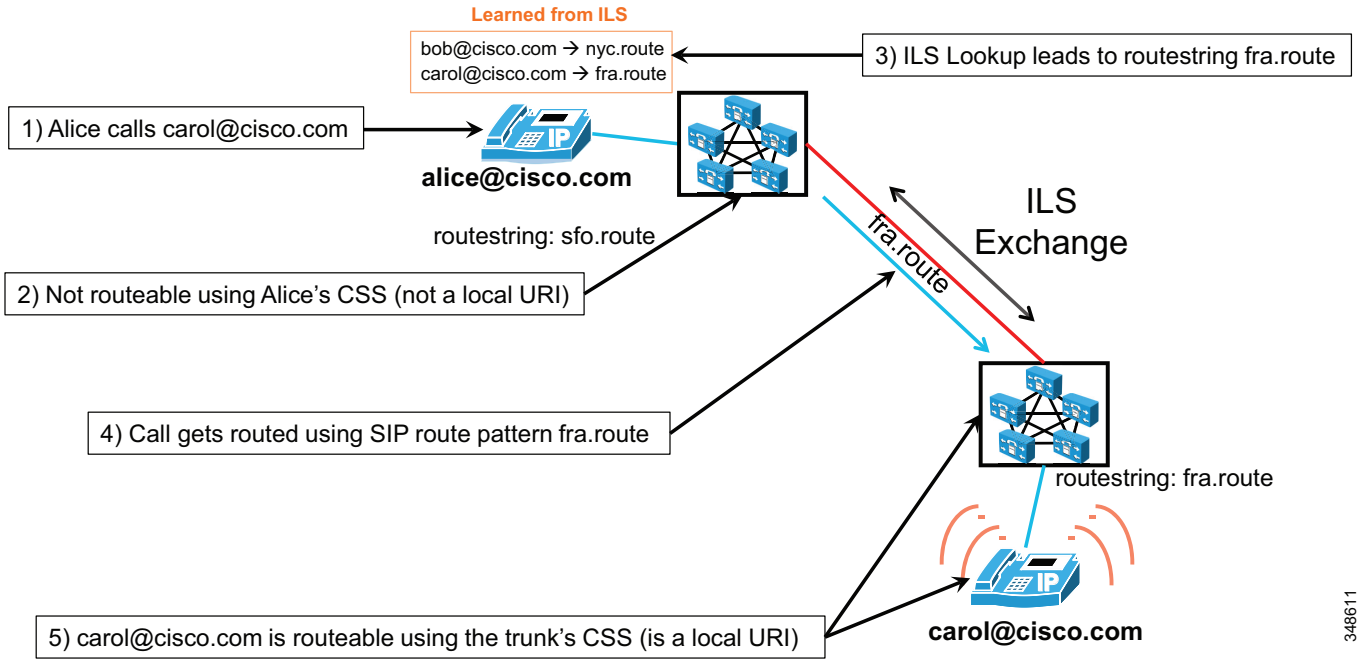


Figure 14-28 shows an example of how a dialed directory URI might be routed by Unified CM. In this example the bottom Unified CM cluster advertises the local directory URI carol@cisco.com. All local directory URIs of this Unified CM cluster are advertised under the SIP routestring fra.route. As part of this information exchange through GDPR, the Unified CM cluster at the top populated its learned directory URI table with the association of carol@cisco.com to the SIP routestring fra.route. If someone then places a call from the phone registered in the top cluster to directory URI carol@cisco.com, the local lookup of directory URI carol@cisco.com will fail because carol@cisco.com is not a local directory URI. The next step in the routing process is to search for carol@cisco.com in the table of directory URIs learned through GDPR. This search will find the information learned from the bottom cluster, and the originating cluster at the top then takes the learned SIP routestring fra.route and tries to find a route by matching this SIP routestring fra.route against the configured SIP route patterns addressed by the calling device's calling search space. A SIP route pattern fra.route is configured and points to a route list that ultimately leads to the SIP trunk pointing to the target Unified CM cluster. The originating Unified CM cluster thus routes the call down to the destination Unified CM cluster. The destination in the sent SIP request will be carol@cisco.com. On the destination cluster, the same routing logic as shown in Figure 14-26 then tries to match carol@cisco.com against all local directory URIs on the destination cluster, which leads to a full match and the target device rings.

The above example shows that the SIP route string namespace is completely independent of the directory URI namespace. There is no requirement to use SIP route strings that are related in any way to the structure of the namespace used for the host portion of directory URIs. This allows to optimize the SIP route string namespace based on the desired routing topology. To disambiguate between SIP route patterns used to directly match on the URI host portion and SIP route patterns used to route directory URIs based on SIP route strings, Cisco highly recommends using an independent namespace for SIP route string route patterns (for example, ".route" or ".ils").

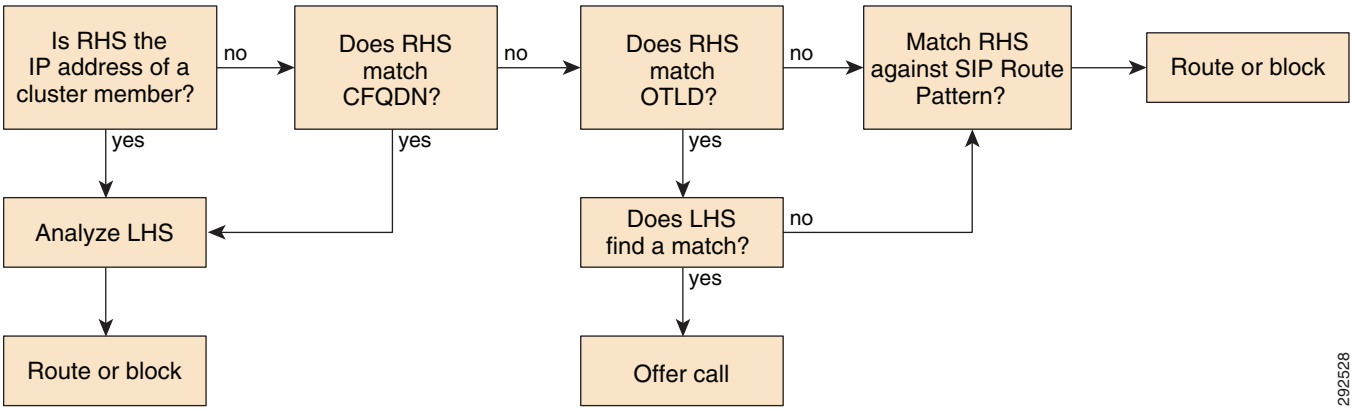
In the above example, the SIP route strings chosen basically identify the individual call controls (fra.route, nyc.route), and the SIP route pattern grid used to route directory URI SIP requests based on learned SIP route strings uses explicit patterns (fra.route, nyc.route) to create the desired reachability. In

a hierarchical topology, hierarchical SIP route strings (for example, sjc.us.route, nyc.us.route, fra.de.route, and muc.de.route) might be used together with wildcard SIP route patterns (*.de.route, *.us.route) routing to the respective aggregating Cisco Unified Communications Manager Session Management Edition (SME) clusters responsible for the addressed set of Unified CM clusters.

Routing Numeric URIs

If a SIP URI is considered to be a numeric URI (see Figure 14-26), the call is handled according to the flowchart shown in Figure 14-29. For Unified CM prior to release 9.0, this is the standard routing procedure for routing of SIP requests.

Figure 14-29 Call Routing Logic for numeric SIP Request



The first step is to check whether the right-hand side of the SIP URI is an IP address or hostname of any server that is a member of the Unified CM cluster or matches the Cluster Fully Qualified Domain Name (CFQDN) configured in Unified CM enterprise parameters. In this case the left-hand side of the URI is considered to be a local numeric pattern and will be matched against numeric patterns existing in local digit analysis using the calling device's calling search space.

The next step is to check whether the right-hand side of the SIP URI matches the Organization Top Level Domain (OTLD) configured in Unified CM enterprise parameters. If this is the case, again Unified CM will try to route the call numerically using the calling device's calling search space. But if no match can be found, then routing will fall back to route the call by matching the right-hand side of the SIP URI against the configured SIP route patterns.

Assuming a Unified CM cluster with cluster members having IP addresses 192.168.10.10, 192.168.10.11, 192.168.20.10, and 192.168.20.11, cluster fully qualified domain name configured as ucm1.cisco.com, and organization top-level domain configured as cisco.com, then all of the following SIP URIs would be routed to local directory number 1234:

- 1234@192.168.10.10
- 1234@192.168.10.11
- 1234@192.168.20.10
- 1234@192.168.20.11
- 1234@ucm1.cisco.com
- 1234@cisco.com

Assuming that no local match for 1234 exists, the first five calls would fail immediately while Unified CM would try to route the sixth call by matching cisco.com against the configured SIP route patterns.

Numeric matching can result in a match on any type of numeric pattern existing locally. This does not only include directory numbers and route patterns and other regular numeric patterns, but can also lead to a match on any numeric pattern learned through GDPR (+E.164 number or pattern, or enterprise number or pattern). If a GDPR learned destination is matched, this immediately leads to a secondary lookup matching the SIP route string of the matched GDPR information against configured SIP route patterns. For the secondary lookup to match the SIP route string, the same calling search space is used that also has been used for the initial numeric lookup. This behavior can be used to restrict access to information learned as part of certain GDPR catalogs by defining a CSS that does not provide access to the SIP route pattern routing the associated SIP route strings.

Note To be able to reach destinations learned through GDPR, the calling device's calling search space has to include the partition that the GDPR learned pattern is residing in and also the partition that the SIP route pattern resides in, which matches the SIP route string associated with the GDPR learned destination.

Cisco TelePresence Video Communication Server

This section provides a high-level overview of the call routing mechanisms available in the Cisco TelePresence Video Communication Server (VCS). For more detailed descriptions, refer to the *Cisco TelePresence Video Communication Server Administrator Guide* and the Cisco VCS deployment guides available at

https://www.cisco.com/en/US/products/ps11337/tsd_products_support_series_home.html

This section covers the following topics:

- Cisco VCS Addressing Schemes: SIP URI, H.323 ID, and E.164 Alias, page 14-53
- Cisco VCS Addressing Zones, page 14-54
- Cisco VCS Pattern Matching, page 14-54
- Cisco VCS Routing Process, page 14-55

Cisco VCS Addressing Schemes: SIP URI, H.323 ID, and E.164 Alias

The Cisco TelePresence Video Communication Server (VCS) enables communications using H.323 and SIP, and it allows any addressing scheme inherently supported by these protocols.

The dialable address formats are:

- IPv4/IPv6 address

  Endpoints and multipoint devices can be called using IP addressing, either IPv4 or IPv6.

- H.323 ID

  The H.323 ID is an alphanumeric identifier for H.323 endpoints. It can be any string of alphanumeric characters. Where SIP and H.323 registration is required for endpoints (dual registration), this alias usually matches the SIP URI.

- E.164 alias

  E.164 uses the same numbering scheme as the PSTN. It is an option that can be configured in H.323 (numbering plan used in the PSTN) together with the H.323 ID.

- SIP URI

  This is an alias that always takes the form *username@domain.*

- ENUM

  ENUM dialing allows an endpoint to be contacted by a caller dialing an E.164 number (a telephone number) even if that endpoint has registered using a different format of alias.

In principle, any SIP URI can be made using E.164 aliases. The username portion of the alias will be the E.164 number, and the hostname portion will be the domain. When configuring this kind of E.164 mapping using SIP, the alias loses information about the user. In this case, FindMe can be configured with the proper alias *username@domain*, thus hiding the complexity of many different addressing schemes. The FindMe alias can be associated to any dialable device, regardless of its addressing scheme.

Cisco VCS Addressing Zones

The VCS receives calls from locally registered endpoints, from neighboring systems, and from endpoints on the public Internet.

Endpoints, gateways, multipoint devices, and content servers registered to the VCS are said to be part of the Local Zone. The Local Zone is further divided into subzones, some of which exist by default and some others might be configured by the administrator.

More generally, a zone is a collection of endpoints that share the same dialing behavior and bandwidth settings. Zones can be local to the VCS or remote.

If dialable entities are not registered on VCS, they might be available on remote zones managed by other call control or systems. These remote zones include: Neighbor Zone, Traversal Client and Traversal Server Zones, DNS Zone, and ENUM Zone.

The concept of Neighbor Zone is analogous to that of a trunk on Cisco Unified CM; it is a SIP or H.323 trunk-side connection to another VCS or Unified CM server or cluster, a third-party call control system, a multipoint device, or a gateway.

A DNS Zone is a non-local destination that can be found using DNS services (SRV). Traversal Client and Server are zones that give access to communications over the Internet using VCS Control and VCS Expressway. An ENUM zone is a non-local destination that is reachable using ENUM services.

Cisco VCS Pattern Matching

Important concepts on VCS routing logic are transforms, also called pre-search transforms, and search rules, also known as searches. The difference between transforms and searches is that searches have a destination target zone, while transforms are configured at system level and cannot be applied per single zone.

Searches and transforms are applied following a priority order configured by an administrator, and they use regular expressions for pattern analysis and string manipulation.

The pre-search transform concept is analogous to translation patterns on Unified CM, with the exception that the use of regular expressions enables alphanumeric transformations.

Search rules are analogous to route patterns in Unified CM. While route patterns are applied to the trunk or route list, search rules have a destination zone as a target.

Both search rules and transforms have the following main characteristics:

- A priority order, which defines the sequential order the VCS uses to analyze the rules or transforms
- A matching expression (pattern string) against which the dialed pattern is checked
- A replacement string, which is the expression used to derive the destination alias

Even though regular expressions allow for complex string manipulation, there are some very common simple applications. One of the most common string manipulations on VCS occurs by adding or stripping the domain part of an alias. An example of this is the following:

Alias: 88302

Search rule matching expression (using a regular expression): (\d+)

Search rule replacement string: \1@cisco.com

Following this simple rule, any dialed number arriving at the VCS will be translated into number@domain. In this case, 88302 will be translated into 88302@cisco.com.

Search rules have the following additional characteristics that can be useful when creating a dialing scheme:

- A target zone (mandatory). The target zone could be the Local Zone for VCS internal calls, or any other zone as a neighbor, traversal client or server, or DNS zone. It might include a policy server as well. The destination zone is selected based on the user's dialed pattern.
- A source zone (optional). Starting with Cisco VCS release 7.2, it is possible to apply a rule only to endpoints calling from a specific zone or subzone.
- A configurable behavior on a successful match of the search rule (mandatory).

On VCS, there is a difference between an alias matching a pattern and an alias that is found and that addresses a device able to answer the call.

If an alias is checked against a search rule matching expression, and the expression matches the alias, VCS will check if the alias exists in the target zone.

If the search rule matches the alias and the alias is found, the call is sent to the target zone.

If the search rule matches the alias but the alias is not found, this means that it does not exist in the target zone. In this case the behavior of VCS depends on what is configured for the **On Successful Match** field of the search rule. If this field is set to **stop**, the routing engine stops even if the alias is not found, and the call is sent to the destination zone. If the field is set to **continue**, the searching process goes on analyzing the remaining lower priority rules until the alias is found, until a rule matches the alias with the **On Successful Match** field set to **stop**, or until all the rules have been analyzed.

This behavior is useful when the administrator does not know where a specific alias is, as in the case of alphanumeric SIP URIs with the same domain registered on multiple call control platforms. As an example, there might be multiple VCSs inside the same company, sharing the same domain company.com. A call for user1@company.com cannot be routed properly if the destination VCS is not known; however, with the routing logic of VCS, it is possible to search for that alias in multiple VCSs or other call control systems and to send the call only after the alias is found.

Cisco VCS Routing Process

When Cisco VCS receives a call, it applies any configured pre-search transforms. After pre-search transforms, the Call Processing Language (CPL) logic applies. Such policies are configured using CPL scripts for advanced routing rules, and might include an external policy server. However, the vast majority of scenarios do not require the use of CPL.

If FindMe aliases are configured, the User Policies are then applied. The FindMe ID is resolved in one or more target aliases, and the call processing logic starts again in order to properly locate the target aliases.

VCS then tries to find a matching expression for the alias by querying the search rules in priority order. If the search rule returns a new destination (SIP URI or alias), the process starts again. This might happen if a call is sent to a DNS or ENUM services, or to a Policy Service.

If the alias is found in any of the zones (Local Zone, neighbor, or so forth) or if a routing destination is returned by the policy service, the VCS will attempt to place the call.

If a match is not found, the VCS will respond with a message to say that the call has failed.

In contrast with Unified CM where the routing logic is based on the longest match, on VCS the logic is priority-based. While changing the order of translation patterns or route patterns on Unified CM does not have any impact in the result of the routing algorithm, changing the rule priority on VCS will lead to different routing behaviors.

Recommended Design

This section provides design guidance and outlines how to implement an end-to-end enterprise dial plan.

Globalized Dial Plan Approach on Unified CM

This section describes dial plan features used to implement simplified call routing based on globalized numbers. The simplification is primarily obtained through the use of a single routing structure for off-net calls, no matter the source of the call. For example, two users in separate countries could use the same route patterns to carry calls to their respective local gateways, instead of requiring site-specific route patterns, each configured to match their respective dialing habits.

The main architectural approach used to attain this globalization can be summarized as follows:

- When a call enters the system, the destination number and the calling number are accepted in their local format but are then immediately globalized by the system. For calls originating from endpoints registered with Unified CM, globalization of the dialed destination achieved through dialing normalization translation patterns and globalization of calling party information is either not required in the case of +E.164 directory numbers or is achieved through appropriate calling party transformation addressed through the phone's calling party transformations for calls from this line. For calls inbound on trunks, inbound called and calling party transformations serve the same purpose.
- Once globalized, the called number is used to route the call to its destination through the use of route patterns expressed in the global form. The global form may be a combination of a global internal, enterprise-specific form such as 81001234 and/or a globalized PSTN representation of a DID number, such as the +E.164 form (for example, +12125551234).
- Once a destination has been identified, the calling and called numbers are localized to the form required by the endpoint, the network, or the system to which the call is to be delivered.

Thus, the guiding principle is:

> Accept localized forms upon call ingress, and globalize them; route the call based on the globalized form; and localize the call to comply to the form required by the destination.

Cisco Unified Communications Manager (Unified CM) offers the following dial plan globalization capabilities:

- Local Route Group, page 14-57
- Support for + Dialing, page 14-57
- Calling Party Number Transformations, page 14-58
- Called Party Number Transformations, page 14-58
- Incoming Calling Party Settings (per Gateway), page 14-59
- Logical Partitioning, page 14-60

Together, these new features enable a Unified CM system to:

- Route calls based on the physical location context of the caller.
- Represent calling and called party numbers in a global form such as that described by the International Telecommunications Union's E.164 recommendation.
- Present calls to users in a format based on local dialing habits.
- Present calls to external networks (for example, the PSTN) in a manner compatible with the local requirements for calling party number, called party number, and their respective numbering types.
- Derive the global form of the calling party number on incoming calls from gateways, based on the calling number digits and the numbering type.
- Control the establishment of calls, as well as the initiation of mid-call features, between endpoints based on policies acting on each endpoint's geolocation, to comply with regulatory requirements in certain countries.

Local Route Group

Local route groups offer the ability to create patterns that route off-net calls to a gateway chosen based on local route group definitions of the originating party. This, for example, allows for egress gateway selection close to the originating party. For example, a single pattern can be defined to route off-net, intra-country calls for all sites within a given country. Phones at every site can be configured to match this pattern, which then would route the call based on the local route group associated with the calling phone and based on the phone's device pool level setting for the respective local route group. This allows a phone in site 1 to route calls through the gateway at site 1, while a phone at site 2, still using the same pattern, would route calls through the gateway at site 2. This feature simplifies the configuration of site-specific routing of off-net calls.

The definition of multiple local route groups allows for differentiated egress gateway selection for different call types so that, for example, different egress gateways can be defined per device pool for emergency, national, and international calls.

Support for + Dialing

Telephone numbers can use the + sign to represent the international dialing access code needed to reach a destination from different countries. For example, +1 408 526 4000 is the international notation for Cisco's main corporate office in the United States. To call this number, an enterprise telephony user from France typically would have to dial 0 00 1 408 526 4000, whereas a caller from the United Kingdom would have to dial 9 00 1 408 526 4000. In each case, + must be replaced with the appropriate off-net access code (as required by the enterprise telephony system) and international access code (as required by the PSTN carrier) relevant for each caller.

The system can route calls directly to destinations defined with +. For example, a user could program a WiFi phone's speed-dial entry for Cisco's main US office as +1 408 526 4000 and dial it directly when roaming in France, the UK, or anywhere else in the enterprise. In each location, the system would translate the destination number into the locally required digit string to allow the call to be routed properly.

Likewise, phone numbers dialed from a dual-mode phone are routable directly over the mobile carrier network when the phone is in GSM mode, or over the enterprise network when the phone is in WiFi mode, if the called number is represented in the +E.164 form. This allows a user to store a single destination number for a particular contact entry, and dial it no matter to which network the phone is currently attached.

This feature allows users to rely on the system to interpret phone numbers represented in the form described by the ITU E.164 recommendation and to route them properly without requiring the user to edit the number to adapt it manually to the local dialing habits.

Calling Party Number Transformations

The calling party number associated with a call routed through Unified CM might sometimes have to be adapted before it is presented to a phone or to the PSTN. For example, a call from +1 408 526 4000 might have to be presented as coming from 408 526 4000 if the destination phone is in the US or Canada, whereas a call from the same number might have to be presented to a destination phone in France as coming from 00 1 408 526 4000. This is mainly to offer users a presentation of the calling party in the customary form offered by their local PSTN, to maintain user familiarity with identification of the origin of calls ringing in.

Calls offered to gateways might require that the calling party number be manipulated to adapt it to the requirements of the telephony carrier to which the gateway is connected. For example, a call from +1 408 526 4000 offered to a gateway located in France might have to represent the calling number as 1 408 526 4000, with a Calling Party Number Type set to International. Similarly, a call from the same number offered to a gateway located in Canada might have to be represent the calling party number as 408 526 4000, with the Calling Party Number Type set to National.

This feature allows the calling party number to be adapted from the form used to route calls within the Unified CM system, to the form required by phone users or off-cluster networks.

Note Some service providers might not be able to accept calling party numbers representing foreign telephone numbers, due to either technical limitations of their equipment, company policies, or governmental regulations. If calling party numbers cannot be accepted by the provider, the provider will either screen and overwrite the calling party number or reject the call. In some networks two calling party identities can exist for a call: user provided and network provided.

Called Party Number Transformations

The called number associated with a call routed through Unified CM might sometimes have to be adapted before it is presented to the PSTN. For example, a call placed to +1 408 526 4000 requires the called party number be transformed to 1 408 562 4000 with the numbering type set to National if it egresses to the PSTN through a gateway located in Canada. If the same call were re-routed toward a French gateway, the called party number would have to be transformed to 1 408 526 4000 with the numbering type set to International.

By manipulating the called party number as well as setting the numbering type for the called number, this feature allows the called party number to be adapted to the form required by off-cluster networks.

At the same time, incoming called party transformations allow the normalization of incoming called party information to a common globalized format before routing the call. Unified CM offers per-gateway settings for this feature, which allow different prefixes for each numbering type to be applied to calls entering different gateways. The settings can be configured on the gateway itself or on the gateway's device pool in order of precedence. A blank entry signifies that no digits will be prefixed; to inherit the settings from the lower-precedence setting, the entry must be set to **Default**. For more complex called party transformations, called party transformation calling search spaces per numbering type can be used. Because SIP does not support the concept of a typed number, for SIP the device pool settings for type Unknown are considered.

Incoming Calling Party Settings (per Gateway)

The calling party number associated with a call as it enters a gateway through a digital interface (for example, ISDN PRI) is also associated with an attribute identifying the calling number's numbering type as either Unknown, Subscriber, National, or International. When combined, the incoming call's calling number and its associated numbering type allow the system to determine the identity of the caller by stripping and prefixing appropriate digits to the incoming call's calling party number. Incoming Calling Party Settings allow the system to apply separate combinations of stripped and/or prefixed digits to the calling party number for each of the four calling number types.

For example, assume two calls come into a gateway located in Hamburg, Germany. Both feature a calling party number of 691234567. The first call is associated with a numbering type of Subscriber. This means the caller is located in Hamburg, thus the city code of Hamburg (40) is implied, as is the country code of Germany (49). Therefore, a full representation of the incoming call is +49 40 69 1234567, which can be obtained by prefixing +49 40 to the incoming call's calling party number for numbering type Subscriber.

The second call is associated with a numbering type of National. This means the caller is located in Germany, and the number already contains the applicable city code (69 is the city code of Frankfurt), but the country code of Germany (49) is implied. A full representation of the second incoming call is thus +49 69 1234567, which can be obtained by prefixing +49 to the second incoming call's calling party number for numbering type National.

Stripping of digits is required to remove digits from the incoming digit string, which must not be part of the globalized number. On some ISDN trunks in Austria, for example, incoming calling party numbers for calls from national destinations have a leading zero which has to be removed to globalize to +E.164. A call from Vienna, for example, could be received with a calling party number of 01666001234 and calling party number type National. For this call the country code of Austria (43) would be implied, and the number already contains the Vienna city code (1). The normalization in this case requires stripping one digit (the leading zero) and prefixing +43 to get to the normalized +E.164 number +43 1 666001234.

Unified CM offers per-gateway settings for this feature, which allow different prefixes for each numbering type to be applied to calls entering different gateways. The settings can be configured on the gateway itself, on the gateway's device pool, or through the cluster-wide service parameters, in order of precedence. A blank entry signifies that no digits will be prefixed; to inherit the settings from the lower-precedence setting, the entry must be set to **Default**.

Due to the global significance of the settings at the service parameter level, Cisco highly recommends using the settings at the device pool level and allowing these settings to be shared for all gateways sharing the same device pool, or at the gateway level if only a single gateway exists that has to use the specific set of transformations. To avoid confusion, Cisco recommends always using only device-pool-level settings or device-level settings in any given installation and not mixing them (using device-level settings for some and device-pool-level settings for others).

For all calls within a given numbering type, the prefix and strip-digits operations are applied, with no consideration for the calling party number originally received.

Note Calls coming from SIP trunks or from SIP gateways are all associated with calling party numbering type Unknown.

In particular, the SIP protocol as implemented on SIP gateways and SIP trunks effectively places the incoming calling party number of all calls in the numbering type Unknown. This prevents Unified CM from applying different calling party number modifications for different calling party number categories.

Unified CM allows the use of Incoming Calling Party Settings Calling Search Spaces (CSSs) for each number type. These CSSs are used to apply modifications to the calling party based on Calling Party Transformation Patterns. These patterns use regular expressions to match a subset of cases, followed by separate digit manipulation operations for each subset. This new capability enables Unified CM to apply different calling party number modifications for different calling party number categories. For example, a SIP trunk used to connect to the PSTN could present calls from local, national, and international parties with the numbering type set to Unknown; then each call's calling party number would be used to match a Calling Party Transformation Pattern in the trunk's CSS associated with number type Unknown, thus allowing Unified CM to apply different calling party number modifications for different calling party number categories.

Logical Partitioning

Some countries such as India have Telecom regulations requiring an enterprise's voice infrastructure to use the local PSTN exclusively when connecting calls outside the enterprise. This requires that the voice system be partitioned logically into two systems: one for Closed User Group (CUG) communications within the enterprise, and a second one to access the local PSTN. A call from an enterprise user in location A to another enterprise user in location B could be made within the CUG system; however, a call from an enterprise user in location A to a PSTN destination, no matter the location, must be made through local access to the PSTN in location A.

While existing dial plan tools can be used to prevent a call from completing if it were placed between endpoints outside the CUG, they are not able to prevent new call legs from being established while the call is in progress. For example, assume that an enterprise user in London, England, calls a co-worker in Delhi, India, over the enterprise network. Once the call is established, the user in Delhi conferences in a customer in India, from the same line on which the call from London was received. This mid-call addition (on the same line) of a destination outside the closed user group is not preventable solely by using the existing dial plan tools in Unified CM (such as Calling Search Spaces and Partitions). Unified CM 7.1 and later releases offer logical partitioning functionality, which allows the establishment and enforcement of policies that apply not only to the initial onset of calls, but also to mid-call features such as conference and transfer.

The combination of globalization features available in Unified CM allows the system to accept calls in the local format preferred by the originating users and carriers, to route the calls on-net using global representations of the called and calling numbers, and to deliver the calls to phones or gateways in the local format required by the destination user or network. These three aspects of the dial plan design approach can be summarized as:

- Localized Call Ingress, page 14-61
- Globalized Call Routing, page 14-62
- Localized Call Egress, page 14-63

Localized Call Ingress

Unified Communications systems with multiple sites located in different regions or countries must satisfy different dialing habits from users and different signaling requirements from the service providers to which gateways are connected. Each local case can be different; this requires that the system be able to "translate" the local dialing habits and signaling requirements into a form that allows for the calls to be routed properly. Therefore, the systems must not only provide for many localized ingress requirements but also yield a single globalized form of any destination pattern.

Localized Call Ingress on Phones

Calls originating on endpoints such as phones or video terminals are typically dialed by users accustomed to a certain set of local dialing habits. Enterprise users in the US are used to dialing 9 1 408 526 4000 to reach Cisco's world headquarters in San Jose, California, whereas users in the UK would dial 9 00 1 408 526 4000 and users in France would dial 0 00 1 408 526 4000. Each of those three dialing forms features an enterprise off-net access code (9 for the US and UK, 0 for France), an international access code (00 for the UK and France, none needed for the US because the destination is intra-country), and a representation of the destination number, including the country code (1). Each of those three groups of users are dialing the same globalized destination number (+1 408 526 4000), but each with their own local dialing habits. In each of the three cases, + can be used as a global abstraction of the local dialing habits.

Unified CM's translation patterns are used to convert localized user input as dialed from phones, to the global form used to route the calls within the Unified Communications system. These patterns must allow all localized dialing habits to be recognized. For details on how these dialing normalizations based on translation patterns can be implemented, see the section on Call Routing in a Globalized Dial Plan, page 14-65.

Phones can also provide dialed strings in the global form of the dialed number. In the case of software endpoints such as Cisco Unified Personal Communicator, + dialing can be accommodated directly from the Telephony User Interface (TUI) of the phone or can be derived from click-to-dial actions taken by the user. On Type-B IP phones, dialing + from the keypad can be achieved by pressing and holding either the * or 0 key, depending on the phone model. Also, the missed and received calls directories can contain entries where the number includes a +. As the user dials from those directories, the resulting call into Unified CM will have a called number beginning with +.

Note For definitions of Type-A and Type-B phones, see Dial Plan Elements, page 14-13.

The calling party number for calls originating from phones is set to the number configured as the directory number of the line from which the call originates. Following the concepts of a globalized dial plan design approach, the calling party information of all calls should be globalized. If the directory number format is not identical to the format chosen for the globalized internal calling party information (+E.164 recommended), then the correct handling of calling party information has to be achieved by properly globalizing the directory number by using the **Caller ID for Calls from this Phone** Calling Party Transformation CSS. This is the recommended way to globalize the calling party information of calls from phones to +E.164, because this method also is compatible with URI-dialed call flows for which calling party transformations in translation patterns are not applicable.

Localized Call Ingress on Gateways

The called and calling numbers delivered into the Unified Communications system by external networks (for example, the PSTN) are typically localized. The form of the numbers may vary, depending on the service provider's configuration of the trunk. As a gateway is connected to a PSTN trunk, the system

administrator must work with the PSTN service provider to determine the applicable signaling rules to be used for this specific trunk. As calls are delivered into the system from the trunk, some of the information about the calling and called numbers will be provided explicitly and some of it will be implied. Using this information, the system must derive the calls' globalized calling and called party numbers.

The globalization of the called party number can be implemented through one of the following methods:

- In the gateway configuration, configure **Call Routing Information > Inbound Calls**, where the quantity of significant digits to be retained from the original called number and the prefix digits to be added to the resulting string are used to globalize the called number. The prefix digits should be used to add the applicable + sign and country, region, and city codes.
- Place translation patterns in partitions referenced by the gateway's calling search space. The translation patterns should be configured to match the called party number form used by the trunks connected to the gateway, and should translate it into the global form. The prefix digits should be used to add the applicable + sign and country, region, and city codes.
- Use the incoming call's called party transformation settings available on the gateway and on the gateway's device pool. There you can define strip and prefix digit instructions or alternatively configure a called party transformation calling search space per numbering type. This is the recommended method.

The globalization of the calling party number should be implemented by using the Incoming Calling Party Settings configured either on the gateway directly or in the device pool controlling the gateway.

Note If the administrator sets the prefix to **Default**, this indicates call processing will use the prefix at the next level setting (device pool or service parameter). Otherwise, the value configured is used as the prefix unless the field is empty, in which case there is no prefix assigned.

For example, assume a call is placed to Cisco's US headquarters (+1 408 526 4000) from a US number, and the call is delivered to a gateway located in San Jose, California. The called number provided to the gateway is 526 4000. This information is sufficient for the Cisco Unified Communications system to derive the full destination number for the call. A call delivered by the service provider on this specific trunk group should inherit an implied country code and area code based on the characteristics of the trunk group connected to the gateway, which presumes that all destination DID numbers handled by the trunk group are from the North American Numbering Plan country code (1) and for area code 408. Therefore, the derived global form of the number is +1 408 526 4000. The calling number provided to the gateway is 555 1234, with the numbering type set to Subscriber. The numbering type allows the system to infer the country code and area code from the configured characteristics of the trunk group. Thus, the system knows that the calling number is +1 408 555 1234.

On a different call, if the calling number is 33158405858 with numbering type International, this is an indication that the global form of the calling number should represented as +33158405858.

Globalized Call Routing

For the destination to be represented in a global form common to all cases, we must adopt a global form of the destination number from which all local forms can be derived. The + sign is the mechanism used by the ITU's E.123 recommendation to represent any global E.164 PSTN number in a global, unique way. This form is sometimes referred to as a fully qualified PSTN number. In this document we refer to this notation as +E.164 (E.164 with leading + sign).

The system can be configured with route patterns that match globalized called numbers including the + sign. These same route patterns can point to route lists and route groups featuring the Standard Local Route Group. This allows for the creation of truly global route patterns because the egress gateway can be determined from the calling endpoint's device pool at the time of the call. All the necessary tasks of adapting the calls (both the calling and the called party numbers) to the local preferences and requirements are performed once a destination has been selected.

Localized Call Egress

When calls are routed to a destination using a global form of the called and the calling numbers, you might have to consider the following localization actions when the call is delivered to its destination.

Phone Calling Party Number Localization

As a call is delivered to a phone, the calling number will be in its global form, which might not be recognizable to the called party. Typically, users prefer to see calls from callers within their country presented with an abbreviated form of the caller's number.

For example, users in the US want to see incoming calls from US callers with a ten-digit national number, without the + sign or the country code (1). If a user whose global phone number is +1 408 555 1234 calls +1 408 526 4000, the called phone would like to receive 408 555 1234 as the calling party number while the phone is ringing. To achieve this, the system administrator should configure a Calling Party Transformation Pattern of: \+1.!, strip pre-dot. The Calling Party Transformation Pattern is placed in a partition included in the destination phone's Calling Party Transformation Pattern CSS, configured at the device-pool level. As a call from +1 408 555 1234 is offered to the phone, it matches the configured Calling Party Transformation Pattern, which removes the +1 and presents a calling party number of 408 555 1234 as the call rings in.

Note On some newer phones, the calling party number stored in the missed and received calls directories is left in its globalized form to allow one-touch dialing from the directories without requiring manual editing of the directories' stored number string. On other phones, the missed and received calls directories store the transformed calling party number. To avoid problems with one-touch dialing from directories, the formats of both the transformed and untransformed calling party number need to match a supported dialing habit. In typical enterprise dial plans this especially precludes localization of calling party information for calls from national numbers to 10 digits because 10-digit-dialing typically cannot be supported as an enterprise dialing habit without creating overlaps with other dialing habits such as abbreviated intra-site dialing.

Note Many phone users are becoming accustomed to the globalized form of PSTN numbers, mainly due to the common use of mobile phones across international boundaries. The system administrator can forgo the configuration of Calling Party Transformation Patterns to localize calling party information for phones if displaying the global form of incoming numbers is preferred.

Gateway Calling Party Number Localization

As a call is delivered to a gateway, the calling party number must be adapted to the requirements of the PSTN service provider providing the trunk group to which the gateway is connected. Calling Party Number Transformation patterns can be used to change the calling party number digit string and numbering type. Typically, a calling party number featuring the gateway's country code should be changed to remove the + sign and the explicit country code, and they should be replaced with the national

prefix. Also, the numbering type of the calling party number should be changed to National. If the gateway is connected to a trunk group featuring a specific area, region, or city code, the specific combination of + sign, country code, and local area code usually must be replaced by the applicable local prefix. Also, the numbering type must be adjusted to Subscriber.

For example, assume that a call from a San Francisco user (+1 415 555 1234) is routed through a route list featuring a San Francisco gateway as a first choice and a Chicago gateway as a second choice. The San Francisco gateway is configured with two Calling Party Transformation Patterns:

- \+1415.XXXXXXX, strip pre-dot, numbering type: subscriber
- \+1.!, strip pre-dot, numbering type: national

As the call is delivered to the San Francisco gateway, the calling party number matches both Calling Party Transformation Patterns. However, the first one is a more precise match and is selected to process the calling party number. Thus, the resulting transformed number is 5551234, with a calling party type set to Subscriber.

If the gateway had not been able to process the call (for example, if all ports were busy), the call would have been sent to the Chicago gateway to egress to the PSTN. The Chicago gateway is configured with the following two Calling Party Transformation Patterns:

- \+1708.XXXXXXX, strip pre-dot, numbering type: subscriber
- \+1.!, strip pre-dot, numbering type: national

As the call is delivered into the Chicago gateway, the calling party number matches only the second Calling Party Transformation Pattern. Therefore, the resulting calling party number offered to the gateway is 4155551234, with a calling party number type set to National.

Gateway Called Party Number Localization

As a call is delivered to a gateway, the called party number must be adapted to the requirements of the PSTN service provider providing the trunk group to which the gateway is connected. Called Party Number Transformation patterns can be used to change the called party number digit string and numbering type. Typically, a called party number featuring the gateway's country code should be changed to remove the + sign and the explicit country code, and they should be replaced with the national prefix. Also, the numbering type of the called party number should be changed to National. If the gateway is connected to a trunk group featuring a specific area, region, or city code, the specific combination of + sign, country code, and local area code usually must be replaced by the applicable local prefix. Also, the numbering type must be adjusted to Subscriber.

For example, assume that a call to a San Francisco user (+1 415 555 2222) is routed through a route list featuring a San Francisco gateway as a first choice and a Chicago gateway as a second choice. The San Francisco gateway is configured with two Called Party Transformation Patterns:

- \+1415.XXXXXXX, strip pre-dot, numbering type: subscriber
- \+1.!, strip pre-dot, numbering type: national

As the call is delivered to the San Francisco gateway, the called party number matches both of the Called Party Transformation Patterns. However, the first one is a more precise match and is selected to process the called party number. Thus, the resulting transformed number is 5552222, with a called party type set to Subscriber.

If the gateway had not been able to process the call (for example, if all ports were busy), the call would have been sent to the Chicago gateway to egress to the PSTN. The Chicago gateway is configured with the following two Called Party Transformation Patterns:

- \+1708.XXXXXXX, strip pre-dot, numbering type: subscriber
- \+1.!, strip pre-dot, numbering type: national

As the call is delivered into the Chicago gateway, the called party number matches only the second Called Party Transformation Pattern. Therefore, the resulting called party number offered to the gateway is 4155552222, with a called party number type set to National.

Note When a call egresses to a gateway, the calling and called party transformation patterns are applied to the calling and called numbers respectively.

Note SIP does not offer an indication of the numbering type. Therefore, SIP gateways are not able to receive an indication of the called or calling party number type set by Unified CM.

Call Routing in a Globalized Dial Plan

The system must be configured to recognize user input and then route and deliver the call to the proper destination. Because the call can originate in many different forms, the system must provide pattern recognition to match each of those forms.

Core routing in the globalized dial plan approach is based on routing +E.164 patterns so that the native dialing habit for this dial plan approach is global +E.164 dialing.

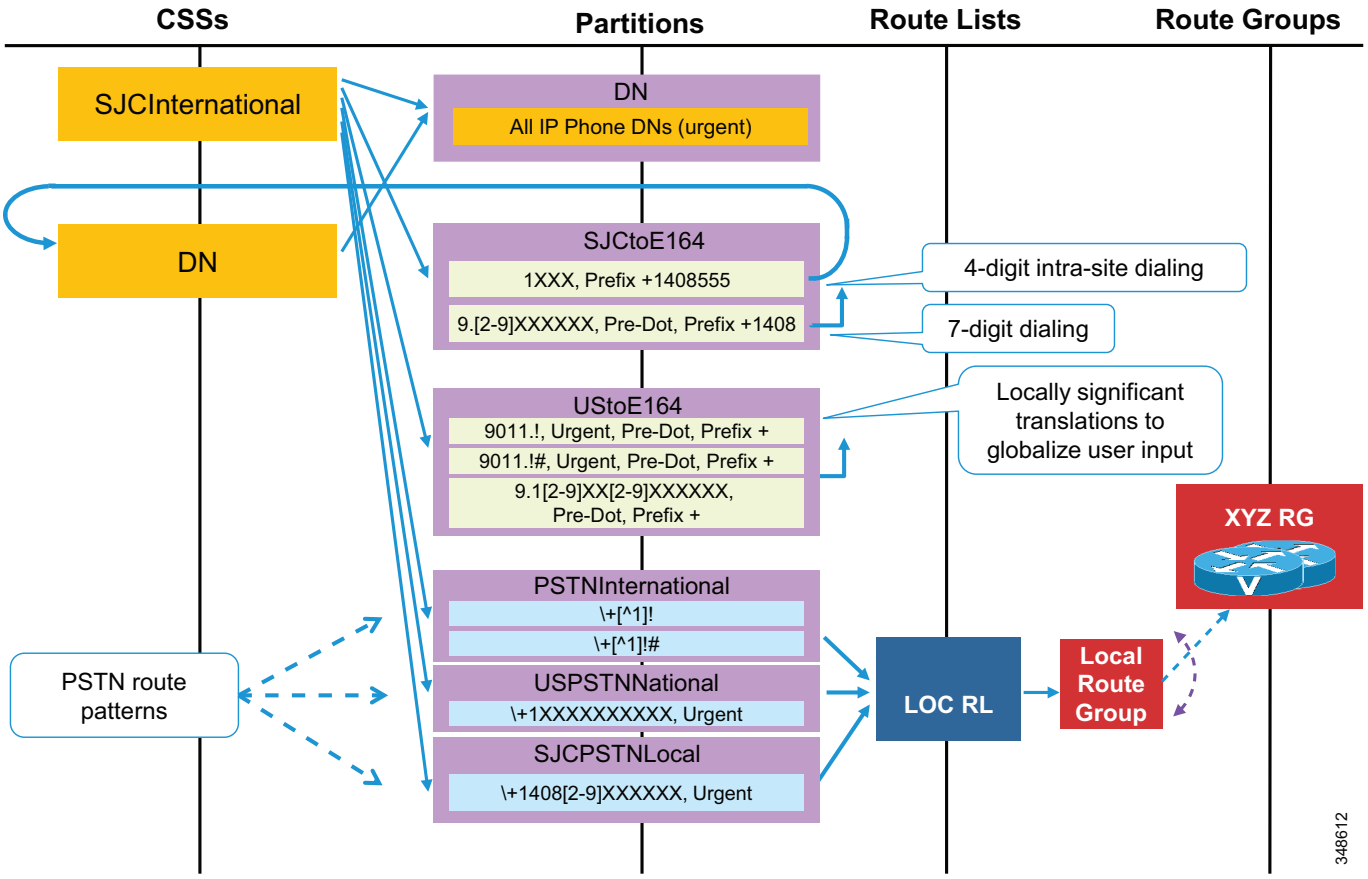
Unified CM's translation patterns are used to convert localized user input as dialed from phones, to the global +E.164 form used to route the calls within the Unified Communications system.

The calling search spaces configured for each site should generally at least allow for:

- Localized intra-site dialing habits of the site
- Localized off-net dialing habits of the users at the site
- Applicable local telephony services such as emergency calls, directory, and operator services
- The globalized form of on-net and off-net numbers

Figure 14-30 shows how to support dialing in the globalized form using local habitual dialing for an example site in the US.

Figure 14-30 Localized and Globalized Dialing



In Figure 14-30, a US IP phone user dials 9011496100773, connects to the destination in Germany, and then releases the call. The called party in Germany calls the US user back, connects, and then releases the call. The US user then goes into the Received calls directory, selects the entry for the last received call (+49 6100 773), and presses Dial.

In this example, the US user initiates two separate calls to the same destination (+496100773). For the first call, the form of the destination number localized for US dialing habits is used, and the corresponding translation pattern 9011.! is matched by the user's input. Once translated, the same calling search space is used for the secondary lookup (**Use Originator's Calling Search Space** set on the translation pattern) and the route pattern \+[^1]! is used to route the call. For the second call, the globalized form of the destination number is used and the route pattern \+[^1]! is used directly.

Comparing these call flows clearly shows the two-step routing process implemented in this dial plan approach: first normalize all dialing habits to +E.164 and then route based on +E.164 patterns. The effective PSTN access level is defined by the PSTN route patterns addressed by the calling search space. More granular access levels can be implemented by adding more specific route patterns.

All directory numbers in partition DN are configured as urgent DNs to avoid potential inter-digit timeout if an on-net destination is called, and the dialed on-net destination overlaps with the variable length off-net route pattern in partition PSTNInternational.

The first translation in partition SJCtoE164 implements 4-digit intra-site dialing, assuming that all local DIDs of the site are in the range +1 408 555 1XXX. Local dialing (9+7) for the site in San Jose is implemented by the second translation pattern in the same partition by again transforming the local habitual dialing to +E.164. The same is true for partition UStoE164, which implements the globalization of US habitual PSTN dialing to international and national destinations.

All dialing normalization translation patterns have **Use Originator's Calling Search Space** set (CSS Inheritance) so that the calling search space used for the secondary lookup, after applying the called party transformations defined in the translation pattern, is identical to the activating calling search space.

The single calling search space creating the requested class of service can be used as a line or device calling search space. In deployments that support mobility features such as extension mobility or device mobility, the line calling search space has to be used to enable the user to keep his class of service when roaming.

A user with extension +1 408 555 1234 can now be reached from other users using the calling search space in the example by dialing:

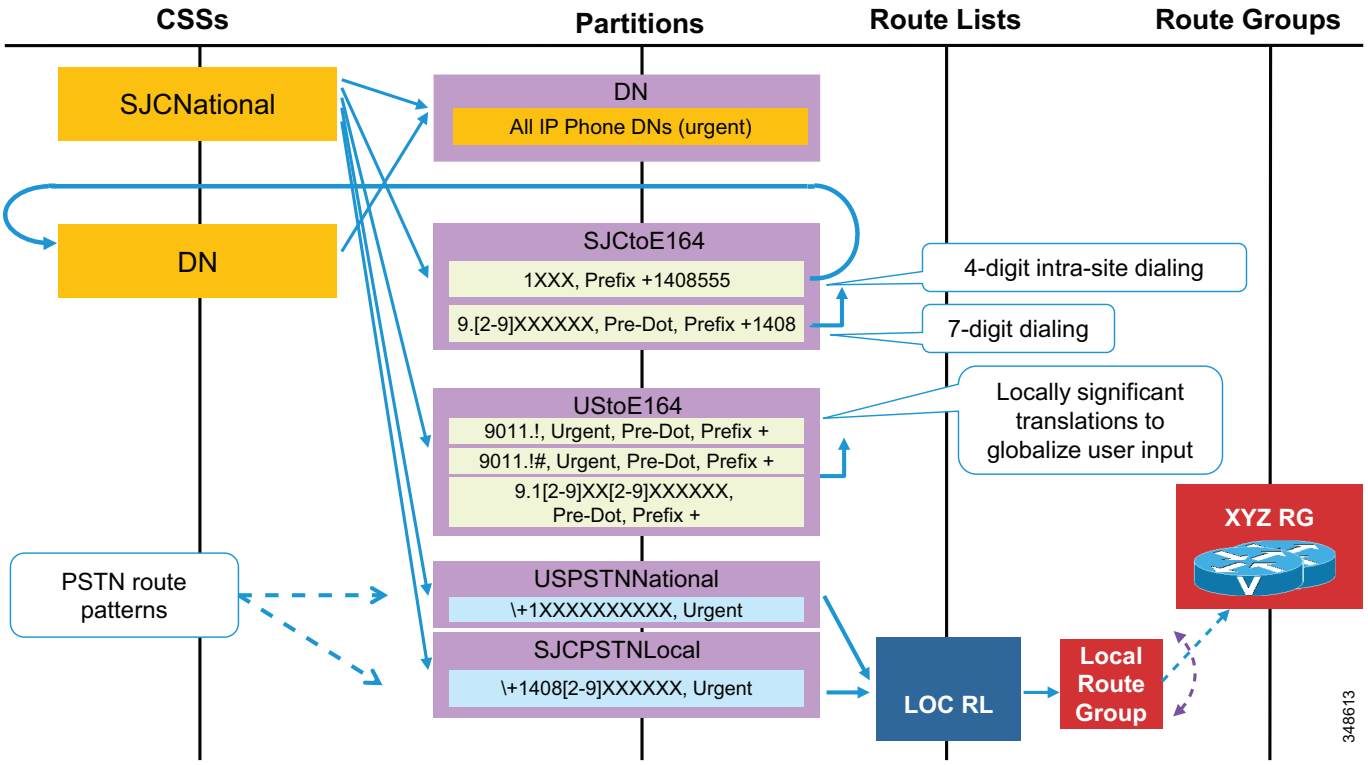
- 1234 — Translation pattern in partition SJCtoE164 transforms dialed digits to +14085551234, and then there is a match on the directory number in partition DN.
- 95551234 — Translation pattern in partition SJCtoE164 globalizes the dialed digits, and then the directory number in partition DN is matched.
- 914085551234 — Translation pattern in partition UStoE164 globalizes the dialed digits, and then the directory number in partition DN is matched.
- +14085551234 — Direct match on the directory number in partition DN.

Other Classes of Service

All translation patterns creating the normalization of dialing habits to +E.164 for class of service "international" in Figure 14-30 use CSS inheritance so that the activating CSS is also used for the secondary lookup after the called party transformations (globalizing to +E.164) are applied. This permits the reuse of the same dialing normalization translation patterns for other classes of service.

Figure 14-31 shows how class of service "national" can be defined based on the same schema used for class of service "international." Comparing this schema to class of service "international" in Figure 14-30, we see that all partitions containing the dialing normalization and PSTN route patterns can be reused. Effectively the only difference is that calling search space SJCNational does not have access to the international PSTN route patterns in partition PSTNInternational.

Figure 14-31 Sharing Dialing Normalization Between Classes of Service



Having access to dialing normalization patterns for the local habitual international dialing 9011 is required even for class of service "national" because we need to support international dialing to international on-net destinations (directory numbers in partition DN outside the US).

More restrictive classes of service such as "local" and "internal" are built following the same schema of simply removing access to the partitions holding the inappropriate PSTN route patterns.

The naming convention used for partitions and calling search spaces in the preceding illustrations helps to identify which pieces of the dial plan need to be replicated to support multiple classes of service, sites, and dialing domains. If the name includes the specification of a site (for example, SJC in partition name SJCtoE164), then that element needs to be replicated for every site. If the name includes the specification of a class of service (for example, International in SJCInternational), then that element needs to be replicated for every class of service. If the name does not include the specification of a site (for example, partition USPSTNNational), then it can be reused for all sites sharing the same dialing habit (in this case all sites in the US).

Calling Unassigned DNs

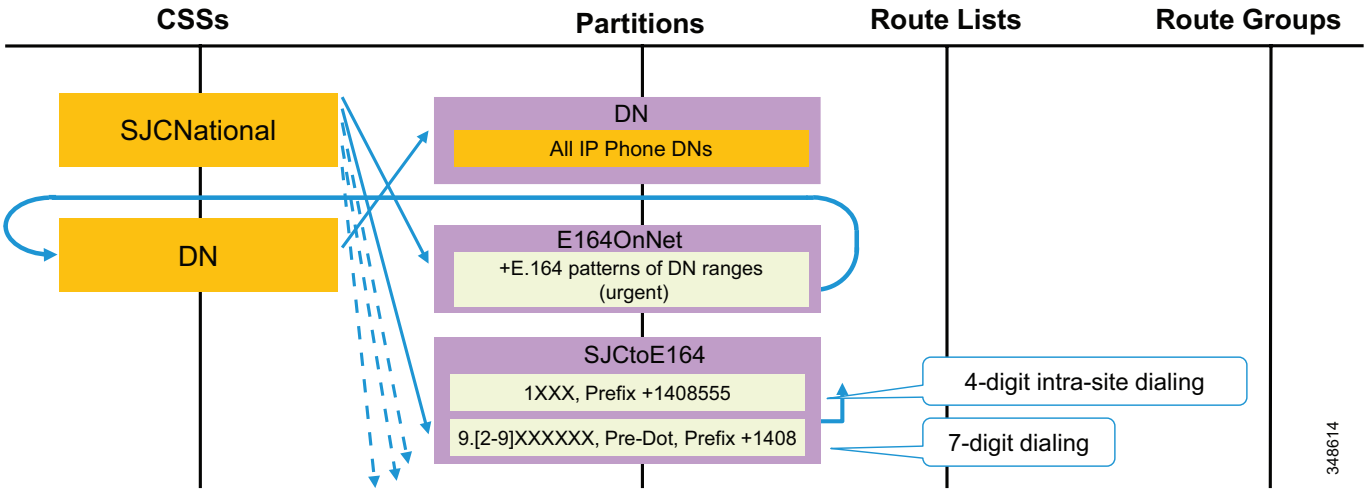
Four-digit dialing normalization translation pattern 1XXX in partition SJCtoE164 does not use CSS inheritance. Instead, this translation pattern uses calling search space DN for the secondary lookup. This is to make sure that if a user dials 1234 and directory number \+14085551234 does not exist, then the call is rejected with cause “unassigned number.” If pattern 1XXX used CSS inheritance, the call would instead be routed to the PSTN after matching the route pattern in partition USPSTNNational. Ultimately this would lead to the same result because the PSTN would either reject the call immediately or route it to the enterprise's PSTN gateway, where it would be seen as an inbound call and then rejected because

the called directory number does not exist. Keep in mind that the inbound calling search space on any gateway typically should have access to internal destinations only and not PSTN destinations, to break routing loops and to avoid toll fraud.

The same PSTN hairpinning problem with calls to unassigned DNs also exists for the other dialing habits implemented by the dialing normalization translation patterns using CSS inheritance and also for +E.164 dialing. For these dialing habits, using the DN calling search space to loop back to the DN partition is not an option because on-net destinations are only a subset of destinations reachable through these dialing habits.

If this hairpinning needs to be avoided, the schema in Figure 14-32 can be used. Here partition E164OnNet holds urgent translation patterns matching the +E.164 prefixes of all on-net destinations. For a site with DID range +1 408 555 1XXX, an urgent translation pattern \+14085551XXX would exist in E164OnNet. These on-net intercept patterns then point back to a DN calling search space that ultimately provides access to the provisioned DNs. All dialing normalization patterns (including the dialing normalization pattern for abbreviated intra-site dialing) use CSS inheritance. A call to an unassigned DN now is not routed to the PSTN because the call is intercepted by the on-net pattern. If the dialed DN then does not exist in the DN partition, the call is rejected with cause “unallocated number.”

Figure 14-32 Avoid PSTN Hairpinning for Unassigned DNs and Support for Non-+E.164 DNs



Another potential purpose of the intercept patterns in E164OnNet is to map from +E.164 to the format of the directory numbers. For example, if the directory numbers are configured as E.164 (without the plus) for a site with DID range +1 408 555 1XXX, a translation pattern \+.14085551XXX with called party transformation "strip pre-dot" (removing the +) would need to be configured in E164OnNet.

Although it is highly recommended to configure directory numbers as +E.164, in some cases the directory numbers might be configured in a different globalized format such as E.164 (without the +), an abbreviated enterprise numbering scheme, or 10 digits in the US. Not configuring directory numbers as +E.164 requires additional number normalization to be configured for globalized caller IDs. Also, some CTI applications (for example, attendant console applications) might require additional number normalization if configured directory numbers do not match the format of numbers stored in global directories.

If +E.164 DNs are used and the rare hairpinning of on-net calls to unassigned DNs is not considered critical, the effort to maintain the list of on-net DN ranges in partition E.164OnNet should be avoided and the simplified dial plan approach shown in Figure 14-30 and Figure 14-31 should be deployed.

Emergency Calls

Access to emergency services has to be granted to all users. This can be achieved either by adding the partition with the emergency number route patterns to each calling search space or by enabling access to the emergency number route patterns through the device-level calling search space. If access to emergency numbers is granted through the device calling search space, then in roaming scenarios (for example, extension mobility) the user has to dial emergency services using the habitual dialing of the visited site, while access to emergency numbers through the line calling search space would allow the user to dial emergency services using the habitual dialing of the home site. This differentiation obviously is important only if the habitual dialing of emergency services differs between home and visiting sites as, for example, in the case of a European user (emergency number 112) logging into an US phone (emergency number 911).

Typically the recommend method is to provide emergency calling services via the emergency number local to the physical location of the calling device. Although this might create overlap between the emergency number and other dialing habits (for example, between 911 and four-digit intra-site dialing starting with 9 for a non-US user from a site with 9XXX abbreviated dialing who logs into a phone in the US), this at least guarantees that any phone in a given location at any time is allowed to place emergency calls using the local habitual emergency dialing independent of whether a remote user from a region with a different emergency number is logged in or not.

To implement this behavior, the emergency patterns needs to be addressed by the device calling search space.

Benefits of the Design Approach

The benefits of the dial plan design approach enabled by the new globalization features include:

- Simplified configuration of call routing, especially when considering local egress to the PSTN
- Simplified configuration and enhanced functionality of system functions such as:
  - Automated Alternate Routing (AAR)
  - Emergency Responder site-specific failover
  - Call Forward Unregistered (CFUR)
  - Tail End Hop Off (TEHO)
  - Click-to-dial of E.164 numbers from soft clients such as Cisco Jabber
  - Adaptive call routing for speed dials originating from roaming extension mobility users or roaming devices
  - One-touch dialing from phone directory entries, including dual-mode phones
  - One-touch dialing from missed and received call lists in IP phone directories

Automated Alternate Routing

If the automated alternate routing (AAR) destination mask is entered in the globalized form, and if every AAR CSS is able to route calls to destinations in the globalized form, then the system administrator can forego the configuration of AAR groups because their sole function is to determine what digits to prefix based on the local requirements of the calling phone's PSTN access to reach the specific destination. A single AAR group with no prefix digits configured can be used for all provisioned devices.

Note The AAR mask or the external phone number mask must be configured in the globalized form on the called destinations to enable AAR, even if the called directory number might already be configured in the globalized form. AAR will be activated only if either the AAR mask or the external phone mask is configured.

Furthermore, in most cases the sole function of the AAR CSS is to route the call to the calling phone's co-located gateway; therefore, it can be configured with only a single route pattern (\+!) pointing to a route list that contains the Standard Local Route Group. Because calls routed by this single route pattern will always be routed through the Local Route Group associated with the calling endpoint, that unique AAR CSS can be used by all phones at all sites, no matter in which region or country they are located.

Cisco Emergency Responder

Call routing to Cisco Emergency Responder is typically implemented by configuring a 911 CTI route point to connect to the primary Emergency Responder server and a 912 CTI route point to connect to the backup Emergency Responder server.

If both Emergency Responder servers are unavailable, 911 calls can be directed to the PSTN egress gateway co-located with the calling phone by configuring:

- The 911 CTI route point to Call Forward No Answer (CFNA) and Call Forward Busy (CFB) to 912, through a calling search space that contains the partition of the 912 CTI route point
- The 912 CTI route point to CFNA and CFB to 911, through a calling search space that contains a global partition, itself containing a route pattern 911 pointing to a route list that contains the Standard Local Route Group

If both CTI route points become unregistered, calls to 911 will be forwarded through the local route group as determined by the calling phone's device pool. If Device Mobility is configured, roaming phones will be associated with the visited site's device pool, and thus associated with the visited site's Local Route Group.

Call Forward Unregistered (CFUR)

To allow calls handled by the Call Forward Unregistered function to use a gateway co-located with the calling phone, configure the CFUR destination of phones using the globalized + form of their PSTN number. The CFUR CSS can be configured with only a single route pattern (\+!) pointing to a route list that contains the Standard Local Route Group. Because calls routed by this single route pattern will always be routed through the Local Route Group associated with the calling endpoint, the same CSS can be used as the CFUR CSS by all phones at all sites, no matter in which region or country they are located.

Tail End Hop Off (TEHO)

To reduce PSTN connectivity charges, system administrators might want to route calls to off-net destinations by using the IP network to bring the egress point to the PSTN as close as possible to the called number. At the same time, if the call's preferred TEHO route is not available, it might be necessary to use the calling phone's local gateway to send the call to the PSTN. This can be achieved by allowing all phones partaking in TEHO routing for a given type of number to match the same route pattern that matches the specific destination number and that points to a route list containing the TEHO egress gateway-of-choice as the first entry and the Standard Local Route Group as the second entry.

Dial Plan with Global Dial Plan Replication (GDPR)

+E.164 alternate numbers and enterprise alternate numbers are defined on a directory number using a mask, and they can be inserted into local digit analysis and advertised to remote call controls. Insertion into local digit analysis and advertising to remote call controls are options that can be enabled independently. A +E.164 or enterprise alternate number needs to be defined only if the number needs to be inserted into local digit analysis, advertised to a remote call control, or used as a PSTN failover number on remote call controls.

Inserting enterprise or +E.164 alternate numbers into local digit analysis effectively creates alternatives to dialing the directory number. For a directory number \+14085551234 in site SJC we could define an enterprise alternate number in partition SJCToE164 using mask 1XXX. This would create a local pattern 1234 in this partition. Using the same enterprise alternate number scheme for all DNs in site SJC would effectively allow removal of the four digit intra-site dialing translation pattern 1XXX shown in Figure 14-30 and Figure 14-31. This schema can be extended to multiple sites because the enterprise alternate numbers that have only local site significance are put into site-specific partitions so that ambiguities are avoided (for example, see Table 14-4).

Table 14-4 Local Site Enterprise Alternate Numbers

Site	DID Range	Enterprise Alternate Number Mask	Enterprise Alternate Number Partition
SJC	+14085551XXX	1XXX	SJCToE164
RTP	+19195552XXX	2XXX	RTPToE164
NYC	+12125551XXX	1XXX	NYCToE164

Using the settings shown in Table 14-4 for directory numbers +14085551234 and +12215551234, the exact same enterprise alternate number 1234 would be created, but both are in different partitions so that the site specificity is preserved.

Although the schema shown in Table 14-4 demonstrates how GDPR enterprise alternate numbers added to local digit analysis can be used to implement abbreviated intra-site dialing without adding dialing normalization translation patterns for this dialing habit, enterprise alternate numbers with only local site significance should never be advertised across GDPR. On the receiving cluster, overlapping (and possibly even identical) enterprise alternate numbers would need to be learned, which causes routing ambiguities.

Note Cisco highly recommends advertising only enterprise alternate numbers with global significance over GDPR. Typically these enterprise alternate numbers follow an enterprise abbreviated on-net numbering plan.

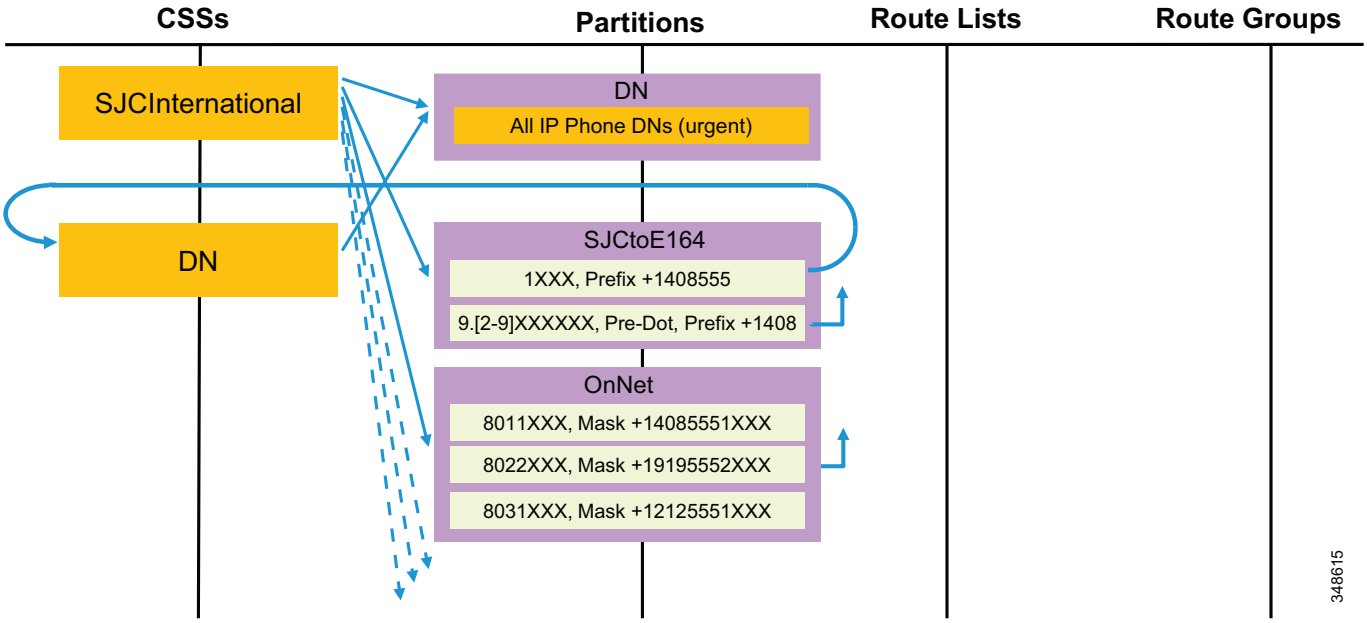
Table 14-5 shows a potential enterprise alternate number schema based on an enterprise abbreviated on-net numbering plan using 8 as the access code and two-digit site numbers.

Table 14-5 Global Enterprise Alternate Numbers

Site	DID Range	Enterprise Alternate Number Mask	Enterprise Alternate Number Partition
SJC	+14085551XXX	8011XXX	DN
RTP	+19195552XXX	8022XXX	DN
NYC	+12125551XXX	8031XXX	DN

These enterprise alternate numbers now have global significance and thus can simply be added into the DN partition implementing the abbreviated inter-site dialing habit for all local directory numbers. The traditional approach to implement the equivalent abbreviated inter-site on-net dialing habit based on dialing normalization translation patterns is shown in Figure 14-33.

Figure 14-33 Abbreviated Intra-Site On-Net Dialing Normalization Translation Patterns



Both schemes (adding enterprise alternate numbers to local digit analysis or using dialing normalization) implement equivalent user experiences. The only difference again is that with dialing normalization patterns, calls to unassigned numbers dialed using this overlay dialing habit are routed to the PSTN and then are hairpinned back. On the other hand, adding explicit enterprise alternate numbers for each directory number to local digit analysis enlarges the local dial plan significantly, which might add complexity to troubleshooting the local dial plan.

Similar to enterprise alternate numbers, +E.164 alternate numbers are also defined by masking the directory number. To define a +E.164 alternate number for a +E.164 DN, the mask simply can be left empty. A +E.164 alternate number of a +E.164 DN should obviously not be added to the local dial plan, but it is still required to be able to advertise the +E.164 alternate number or a +E.164 PSTN failover number to remote call controls.

Using dialing normalization translation patterns to implement abbreviated on-net dialing habits, instead of defining enterprise alternate numbers for each directory number, reduces the complexity of digit analysis because fewer patterns are actually added into the digit analysis. To the same extent, advertising +E.164 and enterprise alternate patterns instead of individual alternate numbers per directory, minimizes the number of advertised dial plan elements and thereby reduces the complexity of dial plans of remote call controls that import the advertised information from GDPR. Advertising only summaries in the form of +E.164 and enterprise patterns is highly recommended.

Call controls that learn dial plan information from GDPR can put the learned information into different partitions based on the type: +E.164 alternate number, enterprise alternate number, +E.164 pattern, and enterprise pattern. If this type-based differentiation is not required to implement the required classes of service, then all numeric dial plan information learned from GDPR can be put into a single partition (such as the OnNet partition in Figure 14-33) that then is added to all calling search spaces implementing classes of service with access to remote on-net destinations.

Differentiated class of service can also be achieved based on limiting access to the SIP route patterns creating the routing schema for the location information in the form of SIP route strings advertised over GDPR. This allows for limiting the reachability of destinations advertised by certain call controls or as part of certain imported GDPR catalogs based on the reachability of the advertised SIP route strings.

Integrating Unified Communications Manager and TelePresence Video Communication Server

Cisco Unified Communications Manager (Unified CM) supports codec registration and alphanumeric URI dialing for the Cisco TelePresence System C Series, EX Series, Profile Series, and SX Series. In this scenario, the Cisco TelePresence Video Communication Server (VCS) can perform two main functions:

- H.323-to-SIP interoperability for video and content
- Business-to-business (B2B) access using VCS Control and VCS Expressway

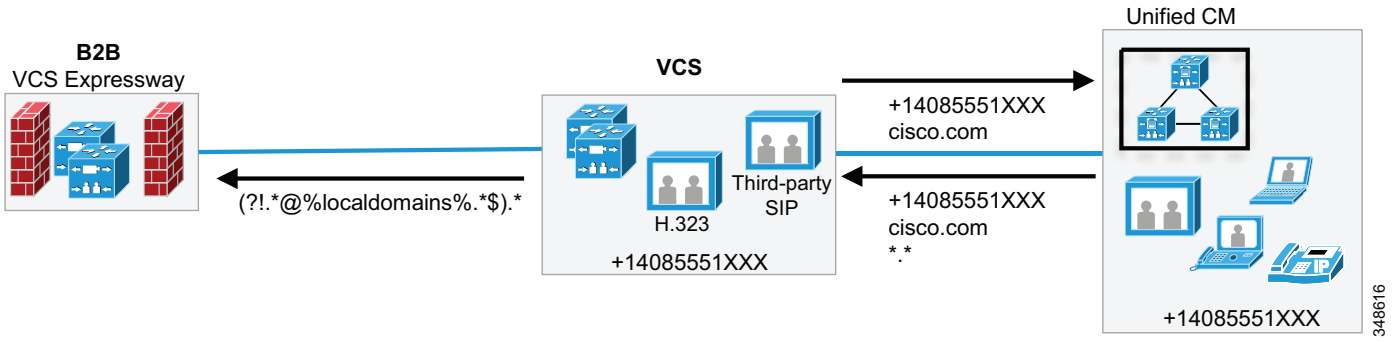
H.323 legacy endpoints can be registered to the VCS, which will perform protocol conversion and content interoperability between H.323/H.239 and SIP Binary Floor Control Protocol (BFCP). Note that VCS behaves as a signaling and media gateway in this scenario, and as such it has to handle the media too, therefore the Interworking feature has to be turned on.

H.323 endpoints connected to the VCS share the same numbering plan used by Unified CM.

Alias manipulation and normalization is done on VCS using the standards-based Portable Operating System Interface for Unix (POSIX) format for regular expression syntax. POSIX is a collection of standards that define some of the matching and replacement functionality that an operating system (UNIX) should support.

Figure 14-34 shows an example topology for interconnecting Cisco VSC and Unified CM to enable end-to-end communications between voice and video endpoints registered with Unified CM and VCS and also communications peers outside the enterprise via VCS Expressway.

Figure 14-34 Sample Topology for Interconnecting Cisco VCS and Unified CM



+E.164 Numbering Plan

To allow for globalized call routing between Unified CM and VCS, a globalized dial plan as described in the section on Globalized Dial Plan Approach on Unified CM, page 14-56, should be implemented on Unified CM and +E.164 addresses need to be provisioned on VCS. Also, the H.323 ID of each endpoint registered with VCS has to be configured with the +E164 number and registered to the Local Zone.

Alias Normalization and Manipulation

Alias normalization has the purpose of presenting the correct alias when dialing an endpoint. Alias normalization might occur at system level or at zone level.

An example of normalization at system level occurs when implementing dial plan transparency with a mixed environment of H.323 and SIP endpoints registered to VCS. H.323 endpoints register the H.323 ID and E.164 alias on the Local Zone of the VCS. However, if a SIP endpoint dials an E.164 alias, it will automatically append the domain, even if the user has just dialed the E.164 number. Regardless of whether the E.164 number is local or remote on another VCS, a normalization rule would strip the domain before forwarding the call. This can be done using a transform.

An example of normalization at zone level occurs when connecting VCS to a Unified CM cluster. In this case Unified CM might use an alias format that does not match the registered endpoint alias. Since this happens only with calls received from Unified CM, a search rule can be applied to normalize the alias before forwarding to the destination.

After normalization, manipulation might also occur. Manipulation after normalization might occur when the call is sent to a non-VCS system that does not support the alias format of VCS.

The following example considers a Unified CM cluster connected to a VCS cluster, where the VCS cluster is used for H.323 endpoint connectivity and B2B connection (see Figure 14-34).

In this case there is no need for alias normalization. All endpoints are reachable through their H.323 ID, which is equal to the +E164 alias. However, calls sent to and received from Unified CM need manipulation before routing them to the final destination.

When a H.323 endpoint calls another H.323 endpoint registered to VCS, it uses the +E.164 number that has been set equal to the H.323 ID (+14085551001 in Figure 14-34), and the call is properly placed.

However when the call is sent to Unified CM, SIP-to-H.323 interworking takes place, and as a consequence a search rule which adds the domain is needed. Assuming that +E.164 numbers from the range +14085551XXX are used on the local VCS, the search rule in Example 14-5 would be required.

Example 14-5 Search Rule for Local +E.164 Destinations on VCS

```
Search Rule "To VCS"
Description: To Local +E164
Priority: 50
mode: alias pattern match
pattern type: regex
pattern string: (\+14085551\d{3})(@.*)
pattern behavior: replace string: \1
On successful Match: Stop
Target: Local Zone
```

Each H.323 client registered to the VCS dialing the internal range would match this rule.

If a +E.164 call comes into VCS from Unified CM, the address dialed would be in any of the following forms:

- +14085551XXX@10.10.10.10:5060 (@ followed by the IP address of the VCS and port number 5060 or 5061 if the trunk configuration on Unified CM uses the IP address of VCS as a peer)
- +14085551XXX@vcs1.cisco.com:5060 (@ followed by the DNS name of VCS and port number 5060 or 5061 if the trunk configuration on Unified CM uses the DNS name of VCS)
- +14085551XXX@cisco.com:5060 (@ followed by the domain and port number 5060 or 5061 if the trunk configuration on Unified CM uses a domain name and a DNS SRV record)

The recommended way to configure the trunk from Unified CM to VCS is to use IP addresses on Unified CM to define VCS as a peer.

The pattern string (\+14085551\d{3})(@.*) in Example 14-5 matches all three of the above formats, and the defined replacement string strips the right-hand side of the received SIP URI to make sure that the received +E.164 address can successfully be matched against the H.323 IDs configured on VCS.

It is possible to use a more stringent pattern matching if a better pattern selection is needed. For example, ([^@]*)@(%ip%|[^@]*cisco.com(.*)). This pattern would match all URIs starting with a sequence of characters that do not include the @, followed by the @ and the IP address of any of the VCS peers in the VCS cluster, or anything that includes "cisco.com" and the port number.

If some SIP endpoints are also registered to VCS, they will automatically add the domain. The search rules above strip the domain even in this case.

For numeric +E.164 calls routed from VCS to Unified CM, a domain has to be added to the SIP URI in the outgoing request because H.323 endpoints do not automatically add a domain. A search rule has to be created in order to add a domain for calls sent to Unified CM, as illustrated in Example 14-6.

Example 14-6 Search Rule "To UCM"

```
Search Rule "To UCM"
Description: To UCM +E164
Priority: 100
mode: alias pattern match
pattern type: regex
pattern string: (\+14085551\d{3})(.*)
pattern behavior: replace string: \1@cisco.com
On successful Match: Stop
Target: UCM Zone
```

The search rule in Example 14-6 makes sure that all numeric dialing from VCS matching \+14085551XXX but not matched by any local client is sent to Unified CM and that the host portion of the SIP URI sent to Unified CM is set to "cisco.com". According to the SIP routing mechanisms of Unified CM as documented in the section on Routing of SIP Requests in Unified CM, page 14-48, and

especially Figure 14-29, the organization top level domain (OTLD) on Unified CM has to be set to "cisco.com" so that Unified CM routes these numeric SIP URIs according to the numeric +E.164 dial plan configured on Unified CM. This rule is also matched by SIP endpoints registered to VCS, if there are any.

To enable B2B connectivity, VCS has to route all B2B calls identified by having a non-local SIP URI host portion to VCS Expressway in the B2B building block. The search rule in Example 14-7 accomplishes this by matching everything that has a domain other than cisco.com and sending it to the VCS Expressway and to the Internet.

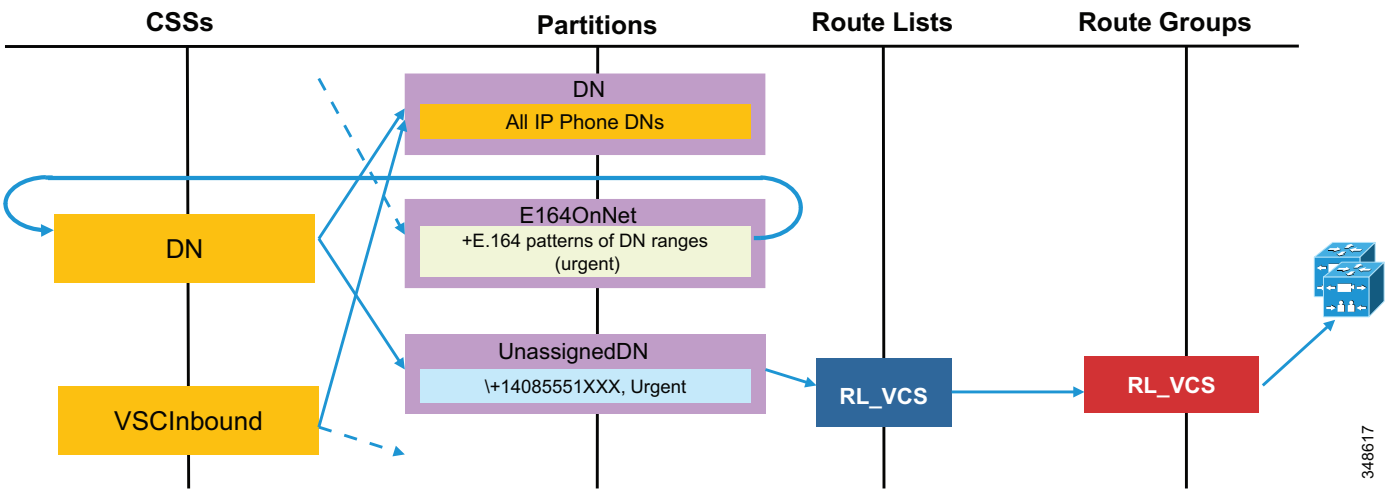
Example 14-7 B2B Search Rule

```
Search Rule "External"
Description: for B2B
Priority: 110
mode: alias pattern match
pattern type: regex
pattern string: [^@]*@[^@]*(?<!cisco.com)
pattern behavior: leave
On successful Match: Stop
Target: VCS-E
```

On Unified CM the +E.164 prefix hosted on VCS has to be added by adding a specific +E.164 route pattern to the Unified CM dial plan and making sure that this route pattern addresses the trunk to VCS by means of an appropriate route list and route group configuration.

If endpoints registered to VCS share the same DN range than the endpoints registered to Unified CM, then the dial plan configuration on Unified CM has to ensure that all +E.164 numbers from the local prefix that are unknown on Unified CM are routed to VCS. Figure 14-35 shows how this can be achieved with a globalized dial plan approach.

Figure 14-35 Intercepting Unassigned Directory Numbers



The globalized dial plan in Figure 14-35 uses the approach as discussed in the section on Globalized Dial Plan Approach on Unified CM, page 14-56. Simply put, the DN calling search space referenced by all dialing normalization translation patterns and the urgent translation patterns matching on the known on-net +E.164 prefixes, has to be extended to include a route pattern matching on the +E.164 prefix that

is shared with VCS. All +E.164 patterns from this range not matched by directory numbers on Unified CM will be matched by this route pattern and sent to VCS. To make sure that no routing loop is created, the inbound calling search space of the trunk coming from VCS should not have access to this route pattern pointing back.

Also, the dial plan on Unified CM has to make sure that all calls dialed as a URI (non-numeric) that do not address a directory URI local to Unified CM, are routed to VCS. The easiest way to achieve this is to add a "catch-all" SIP route pattern (for example, *.*) on Unified CM that also addresses the trunk to VCS through an appropriate route list and route group configuration. Again, to make sure that routing loops are avoided, the inbound calling search space of the trunk coming from VCS should not have access to this "catch-all" SIP route pattern.

Implementing Endpoint SIP URIs

If endpoints on Unified CM can be reached using SIP URIs and +E164 numbers, another search rule can be added to route them properly from VCS to Unified CM, as shown in Example 14-8.

Example 14-8 Search Rule for URI Dialing from VCS to Unified CM

```
Search Rule "URI To UCM"
Description: SIP URI to UCM
Priority: 100
mode: alias pattern match
pattern type: suffix
pattern string: cisco.com
pattern behavior: leave
On successful Match: Stop
Target: UCM Zone
```

If the H.323 endpoints are also addressed using alphanumeric aliases of the same form of a SIP URI instead of using +E.164 aliases, the "To VCS" search rule in Example 14-5 can be replaced by the one in Example 14-9.

Example 14-9 Modified Search Rule "To VCS" Supporting URI Dialing of H.323 Registered Endpoints

```
Search Rule "To VCS"
Description: To Local H.323 aliases
Priority: 50
mode: alias pattern match
pattern type: suffix
pattern string: cisco.com
pattern behavior: leave
On successful Match: Continue
Target: Local Zone
```

"Continue" has to be enabled because, if the alias is not found in the Local Zone, this means that the alias is not local and it will be sent to Unified CM following the next rule of priority 100 ("To CUCM"). However, if the call comes from Unified CM, it will not be sent back to the CUCM zone where the call came from, thus prohibiting routing loops.

On Unified CM a SIP route pattern has to be created to match on "cisco.com" pointing to the same route list used for +E.164 routing to VCS.

If a user on Unified CM dials alice@cisco.com, Unified CM will first match this URI against the locally configured SIP URIs and then as a fallback will match the host portion (cisco.com) against the configured SIP route patterns so that the above SIP route pattern is matched and the call is routed to VCS. If the URI is known on VCS, the call is routed to the endpoint, but the call will not be sent back to Unified CM if the URI is unknown because the call comes from that zone and has not been manipulated.

Special Considerations

This section describes dial plan considerations related to a number of Cisco Unified CM features, including:

- Automated Alternate Routing, page 14-79
- Device Mobility, page 14-83
- Extension Mobility, page 14-84
- Time-of-Day Routing, page 14-91
- Logical Partitioning, page 14-92

Automated Alternate Routing

The automated alternate routing (AAR) feature enables Unified CM to establish an alternate path for the voice media when the preferred path between two endpoints within the same cluster runs out of available bandwidth, as determined by the locations mechanism for call admission control.

The AAR feature applies primarily to deployments with sites connected via a WAN. For instance, if a phone in branch A calls a phone in branch B and the available bandwidth for the WAN link between the branches is insufficient (as computed by the Locations mechanism), AAR can reroute the call through the PSTN. The audio path of the call would be IP-based from the calling phone to its local (branch A) PSTN gateway, TDM-based from that gateway through the PSTN to the branch B gateway, and IP-based from the branch B gateway to the destination IP phone.

AAR can be transparent to the users. You can configure AAR so that users dial only the on-net (for example, four-digit) directory number of the called phone and no additional user input is required to reach the destination through the alternate network (such as the PSTN).

Note AAR does not support CTI route points as the origin or the destination of calls. Also, AAR is incompatible with the Extension Mobility feature when users roam across different sites. Refer to Extension Mobility, page 14-84, for more details.

You must provide the following main elements for AAR to function properly:

- Establish the PSTN Number of the Destination, page 14-80
- Prefix the Required Access Codes, page 14-80
- Select the Proper Dial Plan and Route, page 14-82

Establish the PSTN Number of the Destination

The rerouting of calls requires using a destination number that can be routed through the alternate network (for example, the PSTN). AAR uses the dialed digits to establish the on-cluster destination of the call and then combines them with the called party's AAR Destination Mask; if it is not configured, the External Phone Number Mask is used instead. The combination of the dialed digits and the applicable mask must yield a fully qualified number that can be routed by the alternate network.

Alternatively, by selecting the voicemail checkbox in the AAR configuration, you can allow calls to be directed to the voicemail pilot number. This choice does not rely on the numbers originally dialed by the caller, but routes the call according to the voicemail profile configuration.

Note By default, the directory number configuration retains the AAR leg of the call in the call history, which ensures that the AAR forward to the voice messaging system will select the proper voice mailbox. If you choose "Remove this destination from the call forwarding history," the AAR leg of the call is not present in the call history, which would prevent the automated voice mailbox selection and would offer the caller the generic voicemail greeting.

The AAR Destination Mask is used to allow the destination phone number to be determined independently of the External Phone Number mask. For example, if Caller ID policy for a company required a phone's external phone number mask to be the main directory number of an office (such as 415 555 1000), the AAR destination mask could be set to+1 415 555 1234, to provide AAR with the phone's specific PSTN number.

For example, assume phone A in San Francisco (DN = 2345) dials an on-net DN (1234) configured on phone B located in New York. If locations-based call admission control denies the call, AAR retrieves the AAR Destination Mask of the New York phone (+1212555XXXX) and uses it to derive a number (+12125551234) that can be used to route the call on the PSTN.

It is best to configure the AAR destination mask to yield a fully qualified E.164 number, including the + sign, because this will greatly simplify the overall configuration of AAR. For example, a phone in Paris is configured with an AAR destination mask of +33 1 58 04 58 58. Because this number is a fully qualified E.164 number, it contains all the information required for the Cisco Unified Communications system to derive a routable PSTN number as required by the calling phone's gateway to the PSTN, regardless of whether it is located in France, in Canada, or anywhere else in the world. The following sections elaborate on this approach.

Prefix the Required Access Codes

If the AAR Destination Yields a Fully Qualified E.164 Number Including the + Sign

This is the simplest case; the AAR destination contains + as a wildcard to be replaced by the appropriate access codes require at each gateway. The destination number is ready to be routed to an appropriate route pattern and then transformed at the point of egress to the PSTN by the appropriate called party transformation patterns.

Example 1: A phone in Ottawa, Canada calls a phone in Paris, which triggers AAR due to a lack of bandwidth on the WAN. The AAR destination is +33 1 58 04 58 58. The AAR calling search space of the calling phone contains a route pattern \+!, which routes the call to the Standard Local Route Group. The call is routed to the local gateway in Ottawa, where called party transformation patterns will replace the + with the applicable international access code 011. The resulting call is placed to 011 33 1 58 04 58 58.

Example 2: A phone in Nice, France calls a phone in Paris, which triggers AAR due to a lack of bandwidth on the WAN. The AAR destination is +33 1 58 04 58 58. The AAR calling search space of the calling phone contains a route pattern \+!, which routes the call to the Standard Local Route Group. The call is routed to the local gateway in Nice, where called party transformation patterns will replace the + 33 with the applicable national access code 0. The resulting call is placed to 01 58 04 58 58.

If the AAR Destination Mask Yields a Number Including the Country Code

The destination number (assumed to include the country code) might require a prefix to be routed properly by the origination branch's dial plan. Furthermore, if the point of origin is located in a different area code or even a different country, then other prefixes such as international dialing access codes (for example, 00 or 011) might be required as part of the dialed string.

When configuring AAR, you place the DNs in AAR groups. For each pair of AAR groups, you can then configure prefix digits to add to the DNs for calls between the two groups, including prefix digits for calls originating and terminating within the same AAR group.

As a general rule, place DNs in the same AAR group if they share the same inter-country dialing structure. For example, all phones in the UK dial numbers outside the UK with 9 as a PSTN access code, followed by 00 for international access; all phones in France and Belgium use 0 as a PSTN access code, followed by 00 for international access; all phones in the NANP use 9 as a PSTN access code, followed by 011 for international access.

This yields the following AAR group configuration:

AAR Group	NANP	Cent_EU	UK
NANP	9	9011	9011
Cent_EU	000	000	000
UK	900	900	9

Example 3: A phone in Ottawa, Canada calls a phone in Paris, which triggers AAR due to a lack of bandwidth on the WAN. The AAR destination is 33 1 58 04 58 58. The AAR group of the calling phone is NANP and that of the destination phone is Cent-EU, thus yielding a prefix of 9011. The AAR calling search space of the calling phone contains a site-specific route pattern 9011!, which routes the call to a route list in Ottawa, stripping the 9. The call is routed to the local gateway in Ottawa. The resulting call is placed to 011 33 1 58 04 58 58.

Example 4: A phone in Brussels, Belgium calls a phone in Paris, which triggers AAR due to a lack of bandwidth on the WAN. The AAR destination is 33 1 58 04 58 58. The AAR group of the calling phone and that of the destination phone is Cent-EU, thus yielding a prefix of 000. The AAR calling search space of the calling phone contains a site-specific route pattern 000!, which routes the call to a route list in Brussels, stripping the leading 0. The call is routed to the local gateway in Brussels. The resulting call is placed to 00 33 1 58 04 58 58.

These examples clearly show the benefit of a +E.164 dial plan where no specific AAR groups need to be configured.

These examples clearly show the benefit of a dial plan with +E.164 directory numbers. No specific AAR groups or PSTN prefixes need to be configured. The dialed on-net destinations are already in a format (+E.164) used by the core routing of the dial plan, so that the dialed directory number can be used directly as the PSTN address for the alternate call.

Voicemail Considerations

AAR can direct calls to voicemail. The voicemail pilot number is usually dialed without the need for an off-net access code (if the voicemail pilot number is a fully qualified on-net number, such as 8 555 1000). When AAR is configured to send calls to voicemail, the AAR group mechanism will still prefix the configured access code(s). This configuration requires the creation of an AAR group to be used by all DNs whose desired AAR destination is voicemail (for example, vmail_aar_grp). Ensure that the configuration for this voicemail AAR group uses no prefix numbers when receiving calls from other AAR group DNs.

For example: Assume that DNs located in sites San Francisco and New York are configured with AAR group NANP, which prefixes 9 to calls made between any two DNs in the group. If a DN in San Francisco is configured to send AAR calls to voicemail (for example, 8 555 1000), a call would be placed to 985551000, which would result in a failed call. Instead, the San Francisco DN should be configured with AAR group vmail. The prefix digits for calls from AAR group NANP to AAR group vmail are <none>, as shown in the following table. The call will be placed successfully to 85551000.

AAR Group	NANP	Cent_EU	UK	vmail
NANP	9	9011	9011	<none>
Cent_EU	000	000	000	<none>
UK	900	900	9	<none>

Note When Device Mobility is not used, the AAR group configuration of a DN remains the same even as the device is moved to different parts of the network. With Device Mobility, the AAR group can be determined dynamically based on where in the network the phone is physically located, as determined by the phone's IP address. See Device Mobility, page 14-83, for more details.

Select the Proper Dial Plan and Route

AAR calls should egress through a gateway within the same location as the calling phone, thus causing the completed dial string to be sent through the origination site's dial plan. To ensure that this is the case, select the appropriate AAR calling search space on the device configuration page in Unified CM Administration. Configure the off-net dial plan entries (for example, route patterns) in the AAR calling search space to point to co-located gateways and to remove the access code before presenting the call to the PSTN.

For example, phones at the San Francisco site can be configured with an AAR calling search space that permits long distance calls dialed as 91-NPA-NXX-XXXX but that delivers them to the San Francisco gateway with the access code (9) stripped.

The AAR calling search space configuration can be greatly simplified if the local route group is used in conjunction with using a fully qualified E.164 address (including the + sign) as the AAR destination. This can be achieved by using either a +E.164 AAR destination mask or +E.164 directory numbers. A single calling search space configured with a single partition, containing a single route pattern \+!, pointing to a single route list featuring the Standard Local Route Group, can be used to route the calls of all phones at all sites in an entire cluster. This relies on the pre-configuration of the appropriate gateway-specific called party transformation patterns to adapt the universal form of the destination number to the localized form required by the service provider networks to which the call is delivered at each site.

Note If you have configured additional route patterns to force on-net internal calls dialed as PSTN calls, ensure that these patterns are not matched by the AAR feature. In a globalized dial plan with +E.164 directory numbers, the partition holding these +E.164 directory numbers must not be part of the AAR calling search space.

Note To avoid denial of re-routed calls due to call admission control, AAR functionality requires the use of a LAN as the IP path between each endpoint and its associated gateway to the PSTN. Therefore, AAR dial plans cannot rely on centralized gateways for PSTN access.

Note When Device Mobility is configured, the AAR calling search space can be determined dynamically based on where in the network the phone is physically located, as determined by the phone's IP address. See Device Mobility, page 14-83, for more details.

Device Mobility

Device Mobility offers functionality designed to enhance the mobility of devices within an IP network. (For example, a phone initially configured for use in San Francisco is physically moved to New York.) Although the device still registers with the same Unified CM cluster, it now will adapt some of its behavior based on the new site where it is located. Those changes are triggered by the IP subnet in which the phone is located.

When roaming, a phone will inherit the parameters associated with the device pool associated with the device's current subnet. From a dial-plan perspective, the functionality of the following five main configuration parameters can be modified due to the physical location of the phone. For these parameters to be modified, the device must be deemed as roaming outside its home physical location but within its home device mobility group.

- Local route group

  the roaming device pool's Local Route Group is used. For example, if a device is roaming from San Francisco to New York, the local route group of the New York device pool is used to route calls to the PSTN whenever a pattern points to a route list invoking the Standard Local Route Group.

- Calling party transformation CSS

  The roaming device pool's calling party transformation CSS is used. This allows a phone to inherit the calling party presentation mode that is customary for the phones of the visited location.

- Device calling search space

  The roaming device pool's Device Mobility calling search space is used instead of the device calling search space configured on the device's configuration page. For example, if a device is roaming from San Francisco to New York, the Device Mobility calling search space of the New York device pool is used as the roaming phone's device calling search space. If you use the line/device approach to classes of service, this approach will establish the path taken for PSTN calls, routing them to the local New York gateway.

- AAR calling search space

  The roaming device pool's AAR calling search space is used instead of the AAR calling search space configured on the device's configuration page. For example, if a device is roaming from San Francisco to New York, the AAR calling search space of the New York device pool is used as the roaming phone's AAR calling search space. This calling search space will establish the path taken for outgoing AAR PSTN calls, routing them to the local New York gateway.

- DN's AAR group

  For incoming AAR calls, the AAR group assigned to a DN is retained, whether or not the DN's host phone is roaming. This ensures that the reachability characteristics established for the AAR destination number are retained.

  For outgoing AAR calls, the calling DN's AAR group uses the roaming device pool's AAR group instead of the AAR group selected on the DN's configuration page. Note that this AAR group will be applied to all DNs on the roaming device. For example, all DNs on a device roaming from New York to Paris (assuming both locations are in the same Device Mobility group) would inherit the AAR group configured for outgoing calls in the Paris device pool. This AAR group would be applied to all DNs on the roaming device and would allow for the appropriate prepending of prefix digits to AAR calls made from DNs on the roaming phone.

Call Forward All When Roaming

When a device is roaming in the same device mobility group, Unified CM uses the Device Mobility CSS to reach the local gateway. If a user sets Call Forward All at the phone, if the CFA CSS is set to None, and if the CFA CSS Activation Policy is set to With Activating Device/Line CSS, then:

- The Device CSS and Line CSS get used as the CFA CSS when the device is in its home location.
- If the device is roaming within the same device mobility group, the Device Mobility CSS from the Roaming Device Pool and the Line CSS get used as the CFA CSS.
- If the device is roaming within a different device mobility group, the Device CSS and Line CSS get used as the CFA CSS.

The section on Device Mobility, page 21-14, explains the details of this feature.

Extension Mobility

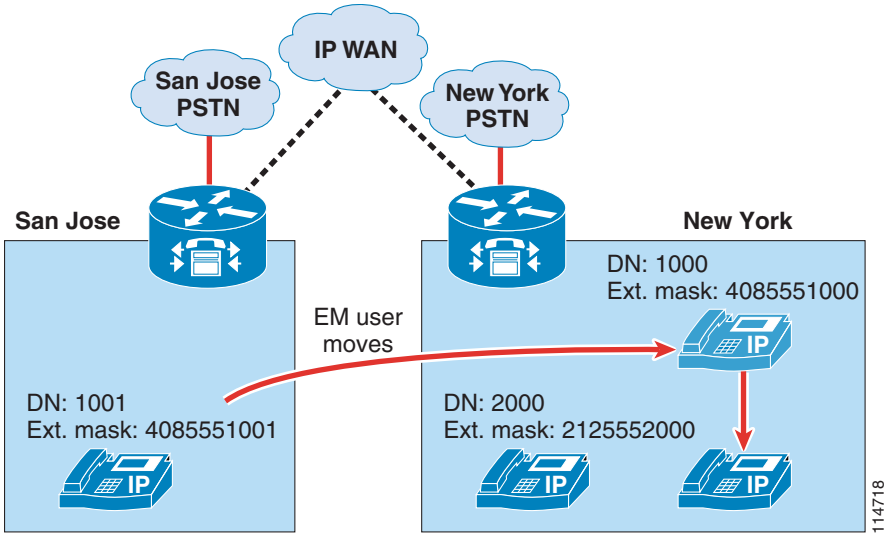
The Extension Mobility feature enables a user to log in to an IP phone and automatically apply his or her profile to that phone, including extension number, speed dials, message waiting indicator (MWI) status, and calling privileges. This mechanism relies on the creation of a device profile associated with each Extension Mobility user. The device profile is effectively a virtual IP phone on which you can configure one or more lines and define calling privileges, speed dials, and so on.

When an IP phone is in the logged-out state, (that is, no Extension Mobility user has logged into it), the phone characteristics are determined by the device configuration page and the line configuration page(s). When a user logs in to an IP phone, the device configuration does not change, but the existing line configuration is saved in the Unified CM database and is replaced by the line configuration of the user's device profile.

One of the key benefits of Extension Mobility is that users can be reached at their own extensions regardless of where they are located, provided that they can log in to an IP phone controlled by the same Unified CM cluster. When Extension Mobility is applied to multisite deployments with centralized call processing, this capability is extended to multiple sites geographically separated from each other.

However, if you combine the Extension Mobility feature with the AAR feature described in the section on Automated Alternate Routing, page 14-79, some limitations exist. Consider the example shown in Figure 14-36, where Extension Mobility and AAR are deployed in a centralized call processing Unified CM cluster with one site in San Jose and one in New York.

Figure 14-36 Extension Mobility and AAR



In this example, assume that an Extension Mobility user who is normally based in San Jose has a DN of 1000 and a DID number of (408) 555-1000. That user's external phone number mask (or AAR mask, if used) is therefore configured as 4085551000. The user now moves to the New York site and logs in. Also, assume that the IP WAN bandwidth between San Jose and New York has been entirely utilized.

When the user in San Jose with extension 1001 tries to call 1000, AAR is triggered and, based on the AAR calling search space of the calling party and the AAR groups of both parties, a new call to 914085551000 is attempted by the San Jose phone. This call uses the San Jose gateway to access the PSTN, but because the DID (408) 555-1000 is owned by that same gateway, the PSTN sends the call back to it. The San Jose gateway tries to complete the call to the phone with extension 1000, which is now in New York. Because no bandwidth is available to New York, the AAR feature is invoked again, and one of the following two scenarios will occur:

- If the gateway's AAR calling search space contains external PSTN route patterns, this is the beginning of a loop that eventually uses all the PSTN trunks at the San Jose site.
- If, on the other hand, the gateway's AAR calling search space contains only internal numbers, the call fails and the caller hears a fast-busy tone. In this case, one PSTN call is placed and one is received, so two PSTN trunks are utilized on the San Jose gateway for the duration of the call setup.

Tip To prevent routing loops such as the one described here, always configure all calling search spaces on the gateway configuration pages to include only internal destinations and no route patterns pointing to route lists or route groups containing that same gateway.

This example highlights the fact that Extension Mobility leverages the dynamic aspect of Cisco IP Communications and, therefore, requires that the call routing between sites use the IP network. Because the E.164 numbers defined in the PSTN are static and the PSTN network is unaware of the movements of the Extension Mobility users, the AAR feature, which relies on the PSTN for call routing, cannot be used to reach Extension Mobility users who move to a site other than their home site.

Note However, if the Extension Mobility user moves to a remote site that belongs to the same AAR group as his or her home site, he or she can use the AAR feature to place calls to other sites when the available IP WAN bandwidth is not sufficient. This is because the path of such a call is determined by the AAR calling search space of the phone from which the call originates. This AAR calling search space does not change when users log in or out of Extension Mobility, and it should be configured to use the visited remote site's gateway.

Tip Configure unregistered Extension Mobility profile DNs to send calls to voicemail. See Call-Forward Calling Search Spaces, page 14-44, for details.

Special Considerations for Cisco Unified Mobility

Cisco Unified Mobility (see the section on Cisco Unified Mobility, page 21-47) relies on functionality that has a direct impact on call routing. To understand the effects of the Cisco Unified Mobility parameters related to dial plans, consider the following example:

Note Only those parameters required in the discussion are mentioned here.

User Paul has an IP phone configured as follows:

- DN: 8 555 1234
- DID number: +1 408 555 1234
- External Phone Number Mask: 408 555 1234
- Line Calling Search Space: P_L_CSS
- Device Calling Search Space: P_D_CSS

Paul's DN is associated with a Remote Destination Profile configured as follows:

- Calling Search Space: P_RDP_CSS
- Rerouting Calling Search Space: P_RDP_Rerouting_CSS
- Calling Party Transformation CSS: P_CPT_CSS

Paul's RDP is associated with a Remote Destination configured as follows:

- Destination Number: +1 514 000 9876 (This is Paul's mobile phone number, on either a single-mode or dual-mode phone.)

Calls from the PSTN placed to Paul or Ringo's DID number are handled by a gateway configured as follows:

- Calling Search Space: GW_CSS
- Significant digits: 7
- Prefix DN: 8

User Ringo has an IP phone configured as follows:

DN: 8 555 0001

DID number: 408 555 0001

External Phone Number Mask: 408 555 0000 (This is the enterprise's main business number.)

Line Calling Search Space: R_L_CSS

Device Calling Search Space: R_D_CSS

The following sections explain the effects of the above mobility parameters on call routing.

Remote Destination Profile

Remote destination profiles (RDPs) are associated with directory numbers (for example, the DN of a user's IP phone) and with remote destinations (for example, the mobile phone number of a user). The RDP controls the interaction between the IP phone and the external numbers (for example, a mobile phone) configured as remote destinations.

Note Remote destinations cannot be configured with on-cluster DNs as destination numbers.

Remote Destination Profile's Rerouting Calling Search Space

When a call is placed to a DN associated with a remote destination profile, the call has the effect of ringing both the DN and the number(s) configured as remote destination(s).

The ability of the caller to reach the destination IP phone is controlled by the caller's Calling Search Space settings. However, the ability for the call to be forked toward the remote destination (for example, a mobile phone) is controlled by the called mobility user's Rerouting Calling Search Space.

For example:
Ringo calls Paul from his IP phone by dialing 8 555 1234. Paul's IP phone rings, as well as his mobile phone.

Here, the ability for Ringo to reach Paul's DN is controlled by the Line and Device calling search spaces on Ringo's IP phone. The dialed destination (8 555 1234) must be in a partition found in the concatenated calling search spaces R_L_CSS and R_D_CSS.

For this same call to be forked to ring Paul's mobile phone, the configured remote destination (+1 514 000 9876) must match a pattern found in the calling search space P_RDP_Rerouting_CSS.

Note Even if the dialing privileges assigned to Ringo's phone do not allow for external calls, the call to the remote destination is handled by the rerouting calling search space associated with Paul's remote destination profile.

Remote Destination Profile's Calling Search Space

A service parameter (Inbound Calling Search Space for Remote Destination) controls which calling search space is used to route calls originating from one of the cluster's remote destinations. Its default setting is Trunk or Gateway Inbound Calling Search Space, which routes all incoming calls using the trunk’s or gateway's configured CSS. If the service parameter is set to Remote Destination Profile + Line

Calling Search Space, then the concatenation of the line CSS of the DN associated with the match's remote destination and the CSS of the Remote Destination Profile associated with the remote destination will be used to route the call.

All the numbers defined as remote destinations within the same cluster will be searched to find a match for any external call coming into the cluster.

The following examples assume that the service parameter Inbound Calling Search Space for Remote Destination is set to Trunk or Gateway Inbound Calling Search Space.

For example:
Paul uses his mobile phone to call Ringo at his desk. The call comes into the gateway from the PSTN, with a calling party number of 514 000 9876 and a called party number of 408 555 0001. The call is routed to Ringo's phone. The number displayed as the calling party number on Ringo's phone is Paul's desk phone number, 8 555 1234. This allows Paul's mobile phone number to remain confidential and allows Ringo's calls placed from the missed and received calls lists to ring into Paul's IP phone, thus making the full set of enterprise mobility features available.

When the call comes into the gateway, the PSTN offers a calling party number of 514 000 9876 and a called party of 408 555 0001. The gateway's configuration will retain the last seven significant digits of the called number and prefix 8, yielding 8 555 0001 as the destination number.

The system detects that the calling party number matches Paul's remote destination number. Upon detecting this match, the system will:

1. Change the calling party number to Paul's DN, 8 555 1234.
2. Route the call to the called number using the incoming gateway's calling search space. Specifically, the routing is done through the GW_CSS calling search space.

The destination (called) number presented by the gateway should be the DN of the phone, and the calling party substitution illustrated in step 1 above renders possible the use of one-touch dialing from the missed/received calls lists.

Note There is no way to partition remote destination numbers. This is worth noting in case multiple user groups (such as different companies, sub-contractors, and so forth) are using the same cluster. When the service parameter Inbound Calling Search Space for Remote Destination is set to Trunk or Gateway Inbound Calling Search Space, the call routing is based on the incoming trunk's or gateway's CSS, regardless of whether or not the calling number matches a remote destination. However, the calling party number substitution still occurs if the calling party matches any remote destination. This means that calls from one tenant's remote destination numbers to another tenant's DID numbers will be presented with a transformed calling party number that matches the caller's on-net extension DN.

Note Any incoming external call where Calling Party Number is not available will be routed according to the incoming gateway's CSS. This also applies to incoming calls from IP trunks, such as SIP or H.323 trunks.

Remote Destination Profile's Calling Party Transformation CSS and Transformation Patterns

Calls originating from an enterprise IP phone to a mobility-enabled DN are forked to both the enterprise destination IP phone's DN and one (or multiple) external destinations. One challenge this creates is to deliver calling party numbers adapted to each destination phone's dial plan. This is to allow for redialing of calls from missed calls and received calls lists. For an enterprise phone, the calling party numbers should be redialable enterprise phone numbers. For a remote destination on the PSTN (such as a home

phone or a mobile phone), the calling party number should be transformed from the enterprise number associated with the calling IP phone to a number redialable from the PSTN (generally, the DID number of the calling phone).

When a call is placed to a mobility-enabled enterprise DN, the associated remote destination profile's calling party transformation calling search space is used to find a match to the caller's calling party number. It contains partitions which themselves contain transformation patterns.

Transformation patterns control the adaptation of calling party numbers from enterprise format to PSTN format. They differ from all other patterns in Unified CM in that they match on the calling number, not the called number. The matching process is done through a regular expression (for example, 8 555 XXXX), and the transformation process allows for the optional use of the calling DN's external phone number mask as well as transformation patterns and digit prefixing.

Once matched, they perform all configured transformations, and the resulting calling party number is used to reach all remote destinations associated with the Remote Destination Profile for which the match occurred.

For example:
When Ringo calls Paul, we want Paul's IP phone to display the calling party number as 8 555 0001 and Paul's mobile phone to display 408 555 0001.

For this case, we create a transformation pattern with the following parameters:

Pattern: 8 555 XXXX

Partition: SJ_Calling_Transform

Use calling party's external phone number mask: un-checked

Calling Party Transformation mask: 555 XXXX

Prefix Digits (outgoing calls): 408

We also have to ensure that partition SJ_Calling_Transform is placed in calling search space P_CPT_CSS.

When the call from Ringo is anchored on Paul's phone, two separate call legs are attempted. The first rings Paul's IP phone and offers the caller's DN as Calling Party Number (that is, 8 555 0001). The second call leg is attempted through Paul's Remote Destination Profile. The RDP's calling party transformation CSS, P_CPT_CSS, is used to find a match for 8 555 0001 in all the referenced partition's transformation patterns. Pattern 8 555 XXXX is matched in partition SJ_Calling_Transform. The transformation mask is applied to the calling party number and yields 555 0001. The prefix digits are added, and the resulting calling party number 408 555 0001 is used when placing the call to the remote destinations.

Note that, in this example, we chose not to use the external phone number mask because it is set to a number different than that of Ringo's DID. This offers flexibility in situations where the calling party number offered to off-net destinations is required to be different based on the relationship of the caller to the called party. The call from Ringo to Paul is between co-workers, thus the disclosure of Ringo's DID number is deemed acceptable. Ringo's next call could be to a customer, in which case the main enterprise number 408 555 0000 is the desired Calling Party Number to be offered to the destination.

Note Calling Party Transformation calling search spaces do not implicitly include the <none> partition; therefore, transformation patterns left in the <none> partition do not apply to any Calling Party Transformation calling search space. This is different from all other patterns in Unified CM, where all patterns left in the <none> partition are implicitly part of every calling search space.

Application Dial Rules

Numbers defined as remote destinations are also used to identify and anchor incoming calls as enterprise mobility calls. Often, the form in which the PSTN identifies calls differs from the form in which an enterprise dial plan requires that calls to external numbers be dialed. Application dial rules can be used to adapt the form in which remote destinations are configured to the form required when forking a call to the remote destination. They allow for the removal from, and prefixing of digits to, the numbers configured as remote destinations.

For example:
Assume the number 514 000 9876 is configured as Paul's remote destination number. This corresponds to the form used by the PSTN to identify calls coming into the enterprise. But it differs from the form used by the enterprise dial plan for outgoing calls, which requires that 91 be prefixed. In this case, we need to create an application dial rule to adapt the remote destination form to the enterprise dial plan's form:

Application Dial Rule:

Name: 514000_ten

Description: Used to prefix 91 to ten-digit numbers beginning with 514000

Number begins with: 514000

Number of Digits: 10

Total digits to be removed: 0

Prefix with Pattern: 91

In this example, calls made from Paul's mobile phone into the enterprise are identified as coming from 514 000 9876. This matches the form in which his number is configured as a remote destination, thus allowing the match to be made and triggering the anchoring of the call on Paul's desk phone as well as adapting the Calling Party Number offered to the on-net destination. (For example, when a call is placed to Ringo's DID number, he sees the call as coming from 8 555 1234.)

When a call is placed to Paul's enterprise DN number, the call leg forked to his remote destination number will be processed by the application dial rule above. The string 514 000 matches the beginning of Paul's remote destination number, and it is ten digits long, so no digits are removed and 91 is prefixed. This yields 91 514 000 9876 as a number to be routed through Paul's Remote Destination Profile calling search space (P_RDP_CSS in this case).

Note This approach offers the ability to reuse calling search spaces already defined to route calls made from IP phones. Creating new calling search spaces not requiring prefixes for outbound calls (that is, ones able to route calls to 514 000 9876 directly) is less preferable because it can create situations where external patterns overlap with on-net patterns.

Time-of-Day Routing

To use this feature, configure the following elements:

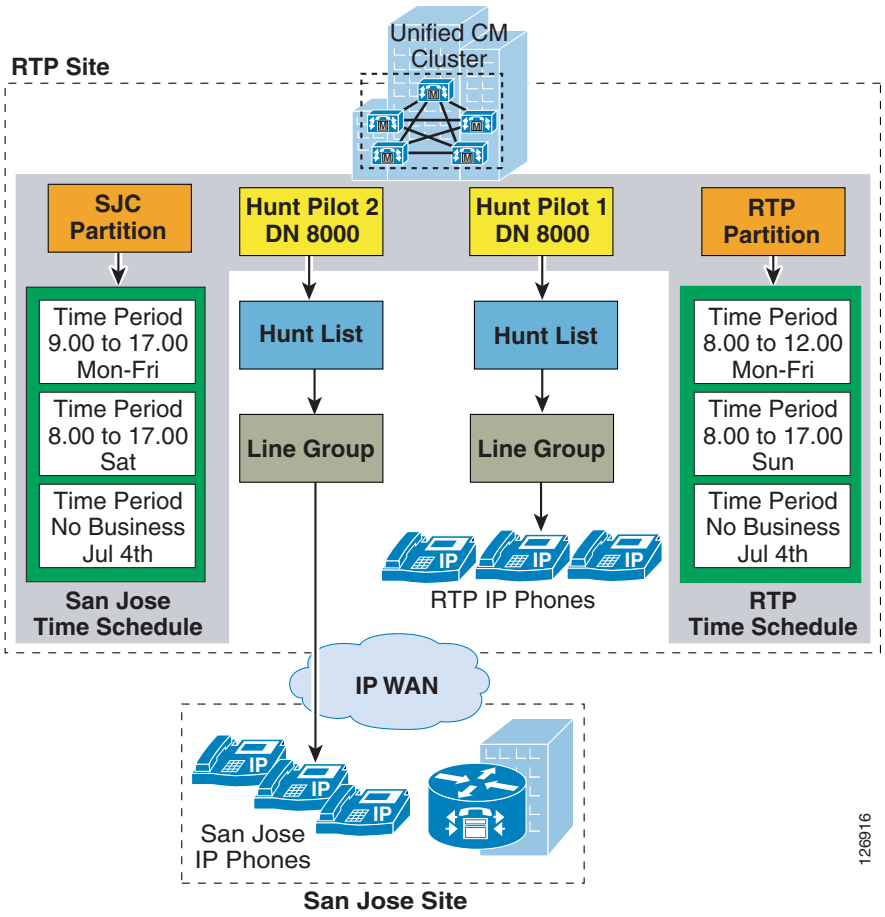
- Time period
- Time schedule

The time period allows you to configure start and end times for business hours. The start and end times indicate the times during which the calls can be routed. In addition to these times, you can set the event to repeat itself on a weekly or yearly basis. Moreover, you can also configure non-business hours by selecting "No business hours" from the Start Time and End Time options. All incoming calls will be blocked when this option is selected.

A time schedule is a group of specific time periods assigned to the partition. It determines whether the partition is active or inactive during the specified time periods. A matching/dialing pattern can be reached only if the partition in which the dialing pattern resides is active.

As illustrated in Figure 14-37, two hunt pilots with the same calling pattern (8000) are configured in two partitions (namely, RTP_Partition and SJC_Partition). Each of these partitions is assigned a time schedule, which contains a list of defined time periods. For example, RTP phones can be reached using Hunt Pilot 1 from 8:00 AM to 12:00 PM EST (GMT - 5.00) Monday through Friday as well as 8:00 AM to 5:00 PM on Sundays. In the same way, SJC phones can be reached using Hunt Pilot 2 from 8:00 AM to 5:00 PM PST (GMT - 8.00) Monday through Friday and 8:00 AM to 5:00 PM on Saturdays. Both of the hunt pilots in this example are inactive on July 4th.

Figure 14-37 Time-of-Day Routing



For the example in Figure 14-37, an incoming call to the hunt pilot (8000) on Wednesday at 3:00 PM will be forwarded to the SJC phones, while a person calling the hunt pilot on July 4th will get a fast busy tone unless there is another pattern that matches 8000.

Logical Partitioning

The elements of logical partitioning include:

- Device types, where phones are classified as *interior*, and gateways and trunks are defined as *border*. Table 14-6 lists the endpoint types for different devices.
- Geolocations, where endpoints are assigned a civic address to be used in policy decisions.
- Geolocation filters, where policy decisions can be made on a subset of the geolocation objects.
- Policies, where communications between endpoints are either allowed or denied based on their comparative (filtered) geolocations and device types.

Note Policies are not applied if all participants in a call (or call attempt) are classified as *interior*. This means that calls between phones on the same cluster are never subjected to logical partitioning policies.

Note Geolocations are not to be confused with locations configured in Unified CM, which are used for call admission control, or with physical locations used for Device Mobility.

Table 14-6 Device Types

Logical Partitioning Device Types	Cisco Unified Communications Manager Device
Border	• Gateway (for example, H.323 gateway) • Intercluster trunk (ICT), both gatekeeper-controlled and non-gatekeeper-controlled • H.225 trunk • SIP trunk • MGCP port (E1, T1, PRI, BRI, FXO)
Interior	• Phones (SCCP, SIP, or third-party) • CTI route points • Analog phones connected to Cisco VG Series Gateways • MGCP port (FXS) • Cisco Unity voicemail (SCCP)

Logical Partitioning Device Types

Unified CM classifies endpoints as either *interior* or *border*. This classification is fixed and cannot be modified by the system administrator.

Geolocation Creation

The (RFC) 4119 standard provides the basis for geolocations. Geolocations use the civic location format specified by the following objects:

- Name
- Description
- Country using the two-letter abbreviation
- State, Region, or Province (A1)
- County or Parish (A2)
- City or Township (A3)
- Borough or City District (A4)
- Neighborhood (A5)
- Street (A6)
- Leading Street Direction, such as N or W (PRD)
- Trailing Street Suffix, such as SW (POD)
- Address Suffix, such as Avenue, Platz (STS)
- Numeric house number (HNO)

- House Number Suffix, such as A, 1/2 (HNS)
- Landmark (LMK)
- Additional Location Information, such as Room Number (LOC)
- Floor (FLR)
- Name of Business or Resident (NAM)
- Zip or Postal Code (PC)

Note In Unified CM, you must define geolocations manually.

Geolocation Assignment

Devices are assigned a geolocation from either the device page, the device pool, or the default Geolocation as configured under Enterprise Parameters, in that order of precedence.

Geolocation Filter Creation

Geolocation filters define which of the geolocation objects should be used when comparing the geolocations of different endpoints. For example, a group of phones may be assigned identical geolocations, except for the room and floor in which they are located. Policies may want to consider endpoints located within the same building as being within the same Closed User Group, and thus allowed to communicate. Even though the actual geolocations of each phone differ, the filtered geolocation is the same. This is useful when policies need to be applied to only the top-level fields of geolocation. For instance, a policy that denies communications between phones and gateways in different cities but allows communications between phones and gateways in the same city, could be based on the comparative filtered geolocations where objects more granular than the City are ignored.

Geolocation Filter Assignment

Phones inherit the filter assignment of their device pool. Gateways and trunks can be configured with a geolocation filter at the device or device pool level, in that order of precedence.

Logical Partitioning Policy Configuration

Logical partitioning policies are configured between geolocation identifiers. A geolocation identifier is the combination of a filtered geolocation and a device type. The filtered geolocation is obtained by taking a device's geolocation and applying the device's associated geolocation filter.

A policy is created as the combination of a set of geolocation objects and a device type (a source geolocation identifier) in relationship with another such combination (the target geolocation identifier). When the relationship is matched, the configured action of "allow" or "deny" is applied to the call leg.

Note Each set of geolocation objects configured in a policy is considered in association with a single device type. For example, a set of geolocation objects such as Country=India, State=Karnataka, City=Bangalore needs to be associated with device type Interior for actions pertaining to Bangalore phones, and separately associated with device type Border for actions pertaining to Bangalore gateways.

Logical Partitioning Policy Application

When user action results in the creation of a new call leg (for example, when a user conferences a third caller into a preexisting call), Unified CM will match the geolocation identifiers of each participant pairs to those of preconfigured policies.

Note When the geolocation identifiers of two devices are being evaluated by logical partitioning, no policy is applied if both devices are of device type Interior. This means that no call, conference, transfer, or so forth, between IP phones within the same cluster will ever be denied due to logical partitioning policies.

For example, consider phones A and B located in Bangalore, India, and gateway C located in Ottawa, Canada. Phone A calls phone B. Because both devices are of type Interior, no policy is invoked. The call is established, and then the user at phone A invokes a conference, which would bring in gateway C. Before the action is allowed, Unified CM will check the geolocation identifiers of A and C, as well as those of B and C, for a match with the preconfigured policies. If any of the matching policies results in a deny action, the new call leg cannot be established.

Note The default policy in Unified CM is deny; in other words, if no policy is configured explicitly to permit a call leg, the call leg will be denied.

In the example above, unless an explicit policy is configured to allow Bangalore Interior devices to connect to Ottawa Border devices, the call leg will be denied.

Emergency Services

Revised: June 14, 2016

Emergency services are of great importance in the proper deployment of a communications system. This chapter presents a summary of the following major design considerations essential to planning for emergency calls:

- 911 Emergency Services Architecture, page 15-2
- Cisco Emergency Responder, page 15-10
- High Availability for Emergency Services, page 15-12
- Capacity Planning for Cisco Emergency Responder Clustering, page 15-13
- Design Considerations for 911 Emergency Services, page 15-13
- Cisco Emergency Responder Deployment Models, page 15-22
- ALI Formats, page 15-29

This chapter presents some information specific to the 911 emergency networks as deployed in Canada and the United States. Many of the concepts discussed here are adaptable to other locales. Please consult with your local telephony network provider for appropriate implementation of emergency call functionality.

In the United States, some states have already enacted legislation covering the 911 functionality required for users in a multi-line telephone system (MLTS). The National Emergency Number Association (NENA) has also produced the *NENA Technical Requirements Document on Model Legislation E9-1-1 for Multi-Line Telephone Systems*, available online at

https://www.nena.org/

This chapter assumes that you are familiar with the generic 911 functionality available to residential PSTN users in North America.

Note The topics discussed in this chapter apply to Cisco Emergency Responder only when it is used in conjunction with Cisco Unified Communications Manager (Unified CM). Cisco TelePresence Video Communication Server (VCS) currently does not support emergency services.

911 Emergency Services Architecture

This section highlights some of the functionality requirements for emergency calls in multi-line telephone systems (MLTS). In the context of this section, emergency calls are 911 calls serviced by the North American public switched telephone network (PSTN).

Any emergency services architecture usually consists of the following elements:

- A distressed caller should be able to dial the emergency services from a fixed line, a mobile phone, a public phone, or any device capable of making the voice call.
- An emergency services call handler must be available to respond to the emergency request and dispatch the needed services such as police, fire, and medical.
- In order to provide help, the call handler should be able to identify the location of the distressed caller as precisely as possible.
- An emergency services network is needed to route the call to the nearest emergency services call handler with jurisdiction for the location of the caller.

The following sections explain some of the important architectural components of 911 emergency services architecture.

Public Safety Answering Point (PSAP)

The public safety answering point (PSAP) is the party responsible for answering the 911 call and arranging the appropriate emergency response, such as sending police, fire, or ambulance teams. The physical location of the phone making the 911 call is the primary factor in determining the appropriate PSAP for answering that call. Generally, each building is serviced by one local PSAP.

To determine the responsible PSAP for a given location, contact a local public safety information service such as the local fire marshal or police department. Also, the phone directory of the local exchange carrier usually lists the agency responsible for servicing 911 calls in a given area.

Typical Situation

- For a given street address, there is only one designated PSAP.
- For a given street address, all 911 calls are routed to the same PSAP.

Exceptional Situation

- The physical size of the campus puts some of the buildings in different PSAP jurisdictions.
- Some of the 911 calls need to be routed to an on-net location (campus security, building security).

Selective Router

The selective router is a node in the emergency services network that determines the appropriate PSAP for call delivery, based on caller's geographic area and the automatic number identification (ANI). The Local Exchange Carrier (LEC) usually operates the selective router. Hence, it is imperative to ensure that the enterprise IP communications network is designed in such a way that the caller is routed to the appropriate selective router based on its location.

Automatic Location Identifier Database

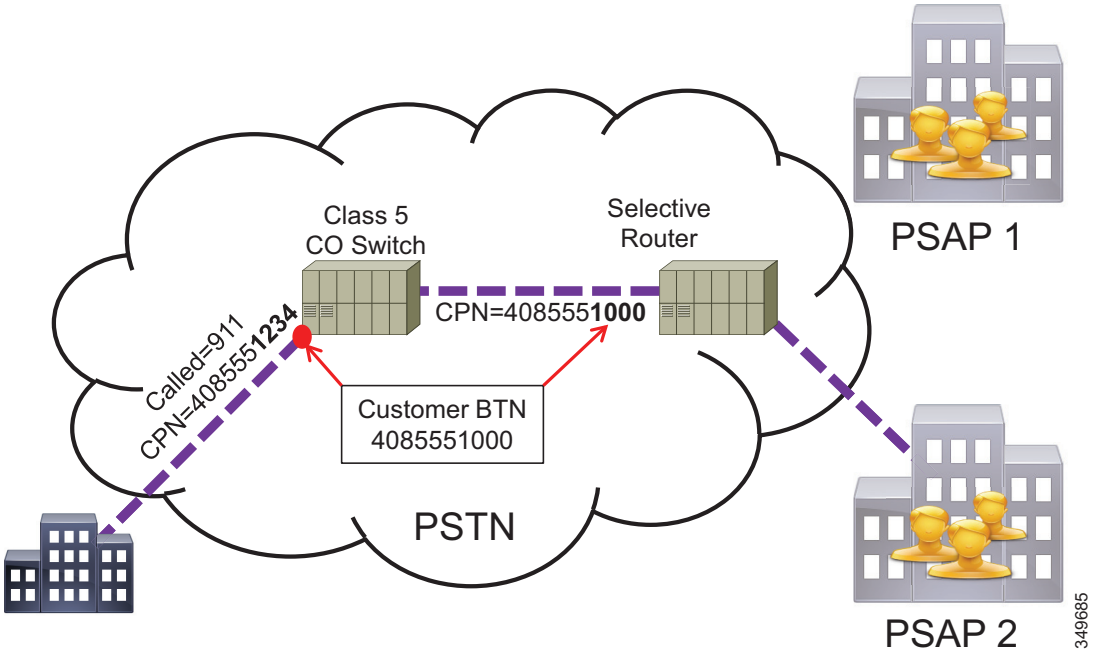
Location information of the caller is an important part of the 911 services infrastructure. The Automatic Location Identifier (ALI) database maintains the location information for the particular geographical location served by the LEC. For every 911 call, the PSAP searches the ALI database to retrieve the caller's location based on the ANI of the calling number. The addresses are stored in the Master Street Address Guide (MSAG) format in the ALI database. The ALI database is maintained on behalf of the local emergency services administration by a contracted third party, generally the incumbent Local Exchange Carrier (LEC).

Service Provider ALI

Service Provider ALI (SP-ALI) refers to a configuration in which the service provider is responsible for defining and maintaining the ALI information for all emergency calls over the connection. SP-ALI service uses the physical interconnection at the LEC to determine the source location of the call. For residential customers, the ALI information is associated with the address of the subscriber and the directory number of that resident. Because the ALI information is determined by the service provider based upon the physical interconnection in the LEC, the subscriber does not have the ability to change or set the ALI information.

The setting of the ALI information based on the physical point of interconnection of the line or trunk applies to PRI trunk connections also. By default, an MLTS operator that uses PRI trunks for PSTN access will have SP-ALI service. The LEC defines the calling party number (CPN) and ALI address for emergency calls. Typically, the calling party number used for emergency calls is the customer's bill-to number (BTN) or the MLTS operator's main number. The physical address associated with the emergency calling number is the address of the demark of the PRI at the customer's facility. If a PRI trunk is set for SP-ALI service, all calls to 911 have the calling party number replaced by the LEC to match the ALI record for the customer. (See Figure 15-1.)

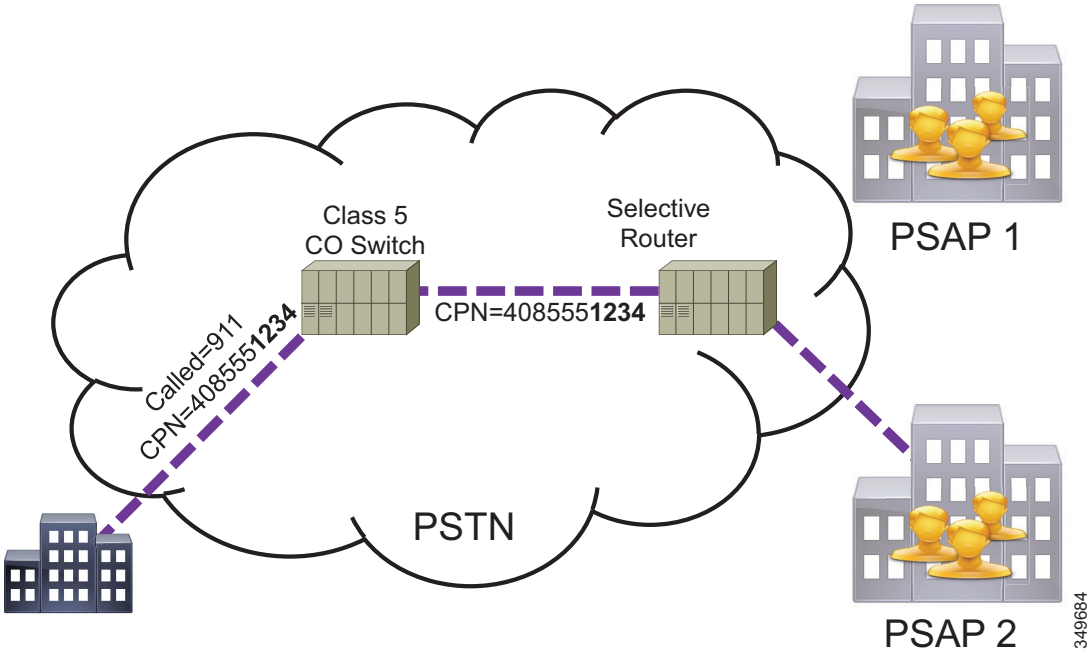
Figure 15-1 Service Provider ALI



Private Switch ALI

Private Switch ALI (PS-ALI) is an enhancement to 911 emergency response systems that enables MLTS operators to provide more specific address and location information for each endpoint. The service allows a customer-generated address table to be loaded into the ALI database so that each station of an MLTS system can be uniquely identified if a call is placed to 911 from that telephone number. The station-specific or location-specific automatic number identification (ANI) generated by the communications system can be passed directly to the E911 system to pinpoint the precise location of the caller. (See Figure 15-2.) The PSAP operator can then direct emergency response personnel to the correct address, building, floor, room, or even cubicle, thereby streamlining operations and increasing accuracy.

Figure 15-2 Private Switch ALI



911 Network Service Provider

After identifying the responsible PSAPs, you must also identify the 911 network service providers to which each PSAP is connected. It is commonly assumed that PSAPs receive 911 phone calls from the PSTN, but that is not the case. Instead, 911 calls are carried over dedicated, regionally significant networks, and each PSAP is connected to one or more such regional networks. In the majority of cases, the incumbent Local Exchange Carrier (LEC) is the 911 network service provider for a PSAP. Some exceptions include military installations, university campuses, federal or state parks, or other locations where the public safety responsibility falls outside the jurisdiction of the local authorities and/or where a private network is operated by an entity other than a public local exchange carrier.

If you are in doubt about the 911 network service provider for a given PSAP, contact the PSAP directly to verify the information.

Typical Situation

- For a given street address, the 911 network service provider is the incumbent Local Exchange Carrier (LEC). For a location served by Phone Company X, the corresponding PSAP is also served by Phone Company X.
- All 911 calls are routed directly to an off-net location, or all 911 calls are routed directly to an on-net location.

Exceptional Situation

- The local exchange carrier (LEC) through which the MLTS interfaces to the PSTN is *not* the same LEC that serves as 911 network service provider to the PSAP. (For example, the communications system is served by Phone Company X, but the PSAP is connected to Phone Company Y.) This situation might require either a special arrangement between the LECs or special, dedicated trunks between the phone system and the PSAP's 911 network service provider.
- Some LECs may not accept 911 calls on their networks. If this is the case, the only two options are to change LECs or to establish trunks (dedicated to 911 call routing) connected to a LEC that can route 911 calls to the appropriate PSAPs.
- Some (or all) of the 911 calls have to be routed to an on-net location such as campus security or building security. This situation can easily be accommodated during the design and implementation phases, but only if the destination of 911 calls for each phone has been properly planned and documented.

Interface Points into the Appropriate 911 Networks

For larger communications systems, 911 connectivity might require many interface points. Typically, more than one E911 selective router is used within a LEC's territory, and these routers usually are *not* interconnected.

For example, an enterprise with a large campus could have the following situation:

- Building A located in San Francisco
- Building B located in San Jose
- San Francisco Police Department and San Jose Police Department are the appropriate PSAPs
- San Francisco Police Department and San Jose Police Department are served by the same 911 network service provider
- However, San Francisco Police Department and San Jose Police Department are served by different E911 selective routers operated by that same 911 network service provider!

This type of situation would require two separate interface points, one per E911 selective router. The information pertaining to the E911 selective router territories is generally kept by the incumbent LEC, and the local account representative for that LEC should be able to provide an enterprise customer with the pertinent information. Many LECs also provide the services of 911 subject matter experts who can consult with their own account representatives on the proper mapping of 911 access services.

Typical Situation

- For single-site deployments or campus deployments, there is usually only one PSAP for 911 calls.
- If access to only one PSAP is required, then only one interface point is required. Even if access to more than one PSAP is required, they might be reachable from the same E911 selective router, through the same centralized interface. If the enterprise's branch sites are linked via a WAN (centralized call processing), it is desirable to give each location its own local (that is, located inside each branch office) access to 911 to prevent 911 isolation during WAN failure conditions where Survivable Remote Site Telephony (SRST) operation is activated.

Exceptional Situation

- The physical size of the campus puts some of the buildings in different PSAP jurisdictions, *and*
- Some of the 911 calls have to be routed to different E911 selective routers, through different interface points.

Note Some of the information required to establish the geographical territories of PSAPs and E911 selective routers is available online or from various competitive local exchange carrier (CLEC) information web sites. (For example, https://clec.att.com/clec/hb/shell.cfm?section=782 provides some valuable data about the territory covered by AT&T in California and Nevada.) However, Cisco strongly recommends that you obtain proper confirmation of the appropriate interface points from the LEC prior to the design and implementation phases of 911 call routing.

Interface Type

In addition to providing voice communications, the interfaces used to present 911 calls to the network must also provide identification data about the calling party.

Automatic Number Identification (ANI) refers to the North American Numbering Plan number of the calling party, which is used by networks to route a 911 call to the proper destination. This number is also used by the PSAP to look up the Automatic Location Identification (ALI) associated with a call.

911 calls are source-routed, which means that they are routed according to the calling number. Even though different locations are all dialing the same number (911), they will reach different PSAPs based on their location of origin, which is represented by the ANI (calling number).

You can implement 911 call functionality with either of the following interface types:

- Dynamic ANI assignment
- Static ANI assignment

While dynamic ANI assignment scales better (because it supports multiple ANIs) and lends itself to all but the smallest of applications, static ANI assignment can be used in a wider variety of environments, from the smallest to the largest systems.

Dynamic ANI (Trunk Connection)

The dynamic aspect of ANI refers to the fact that a communications system has many endpoints sharing access to the 911 network across the same interface, and the ANI transmitted to the network might need to be different for each call.

There are three main types of dynamic ANI interfaces:

- Integrated Services Digital Network Primary Rate Interface (ISDN-PRI, or simply PRI)
- Session Initiation Protocol (SIP) trunk
- Centralized Automatic Message Accounting (CAMA).

PRI

This type of interface usually connects a communications system to a PSTN Class 5 switch. The calling party number (CPN) is used at call setup time to identify the E.164 number of the calling party.

Most LECs treat the CPN differently when a call is made to 911. Depending upon the functionality available in the Class 5 switch and/or upon LEC or government policy, the CPN may not be used as the ANI for 911 call routing. Instead, the network may be programmed to use the listed directory number (LDN) or the bill-to number (BTN) for ANI purposes.

If the CPN is not used for ANI, then 911 calls coming from a PRI interface all look the same to the 911 network because they all have the same ANI, and they are all routed to the same destination (which might not be the appropriate one). The replacement of the CPN by the LEC is typically called Service Provider ALI (SP-ALI), because the service provider specifies the CPN for ALI lookup.

Some LECs offer a feature to provide CPN transparency through a PRI interface for 911 calls. With this feature, the CPN presented to the Class 5 switch at call setup is used as ANI to route the call. The feature name for this functionality varies, depending on the LEC. (For example, SBC calls it Inform 911 in California.)

Note When SP-ALI service is used, the CPN *must* be a routable North American Numbering Plan number, which means that the CPN must be entered in the routing database of the associated E911 selective router.

Note For Direct Inward Dial (DID) phones, the DID number could be used as the ANI for 911 purposes, but only if it is properly associated with an Emergency Service Number in the 911 service provider's network. For non-DID phones, use another number. (See Emergency Location Identification Number Mapping, page 15-14, for more information.)

Many Class 5 switches are connected to E911 selective routers through trunks that do not support more than one area code. In such cases, if PRI is used to carry 911 calls, then the only 911 calls that will be routed properly are those whose CPN (or ANI) have the same Numbering Plan Area (NPA) as the Class 5 switch.

Example

An MLTS is connected to a Class 5 switch in area code 514 (NPA = 514). If the MLTS were to send a 911 call on the PRI trunk, with a CPN of **450**.555.1212, the Class 5 switch would send the call to the E911 selective router with an ANI of **514**.555.1212 (instead of the correct **450**.555.1212), yielding inappropriate routing and ALI lookup.

To use PRI properly as a 911 interface, the system planner must ensure that the CPN will be used for ANI and must properly identify the range of numbers (in the format NPA NXX TNTN) acceptable on the link. For example, if a PRI link is defined to accept ANI numbers within the range 514 XXX XXXX, then only calls that have a Calling Party Number with NPA = 514 will be routed appropriately.

SIP Trunk

SIP trunking is an IP-only interface that connects a communications system to a service provider, typically through a Session Border Controller (SBC). SIP trunks allow for the same dynamic calling party number delivery to the carrier as PRI trunks; but unlike PRI trunks, SIP trunks do not have a physical limit on the number of calls that can be established concurrently.

When emergency services are called over a SIP trunk, delivery of the call to the correct selective router must be verified with the provider. Unlike PRI circuits that terminate at the local LEC, SIP trunks might not have a physical connection with the local LEC and as a result will not automatically route 911 calls to the selective router in the municipality of the calling party. Additionally, each SIP trunk provider might have different E911 routing capability; for example, one service provider may be able to deliver calls to selective routers across the US based upon the calling party number (even outside the local area), while another service provider may allow E911 calls into only one customer-specified selective router. A Cisco Unified CM administrator should always confirm the 911 call delivery capabilities with the carrier, especially when a SIP trunk is providing centralized call routing.

SIP service providers are required to route 911 calls to the appropriate rate center or PSAP for any DID number that they service over a SIP trunk. For example, assume that a deployment has a SIP trunk that physically terminates in a data center in Dallas Texas that services DIDs for a San Francisco office with the range of 415-555-1xxx and for a New York office with the range of 212-448-2xxx. If a call to 911 is place from 415-555-1800, then the SIP provider must route the call to the San Francisco selective router for PSAP delivery. If a user at extension 212-448-2840 in the New York City office dials 911, the call can be routed on the same SIP trunk to the appropriate selective router in the New York City area to reach the PSAP appropriate for the caller.

CAMA

Centralized Automatic Message Accounting (CAMA) trunks also allow the MLTS to send calls to the 911 network, with the following differences from the PRI approach:

- CAMA trunks are connected directly into the E911 selective router. Extra mileage charges may apply to cover the distance between the E911 selective router and the MLTS gateway point.
- CAMA trunks support 911 calls only. The capital and operational expenses associated with the installation and operation of CAMA trunks support 911 traffic only.
- CAMA trunks for the MLTS market may be limited to a fixed area code, and the area code is typically implied (that is, not explicitly sent) in the link protocol. The connection assumes that all calls share the same deterministic area code, therefore only 7 or 8 digits are sent as ANI.

Static ANI (Line Connection)

Static ANI provides a line (rather than a trunk) connection to the PSTN, and the ANI of the line is associated with all 911 calls made on that line, regardless to the CPN of the calling phone. Static ANI is based on the physical interconnection point in the LEC. Because the Static ANI is defined by the carrier on the interconnection point in the LEC, Static ANI emergency call routing is also referred to as Service Provider ALI (SP-ALI). A plain old telephone service (POTS) line is the most common type of connection used for this purpose.

POTS lines are one of the simplest and most widely supported PSTN interfaces. A POTS line usually comes fully configured to accept 911 calls. In addition, the existing E911 infrastructure supports 911 calls from POTS lines very well.

The POTS approach has the following attributes:

- The operational costs associated with a POTS line are low.
- The POTS line can even serve as a backup line in case of power failure.
- The POTS line number can be used as the callback number entered into the ALI database.
- POTS lines represent the lowest cost 911 support for locations where user density does not justify local PRI or CAMA access into the PSTN.
- POTS lines are ubiquitous in PSTN installations.

All outgoing 911 calls through this type of interface are treated the same by the E911 network, and any tools that enable ANI manipulation presented to the E911 network (such as translations or transformations) are irrelevant because the ANI can be only the POTS line’s number.

Cisco Emergency Responder

Ease of administration for moves, adds, and changes is one of the key advantages of IP communications technology. To provide for moves, adds, and changes that automatically update 911 information without user intervention, Cisco has developed a product called the Cisco Emergency Responder (Emergency Responder).

Cisco Emergency Responder provides the following primary functionality:

- Dynamic association of a phone to an Emergency Response Location (ERL), based on the detected physical location of the phone.
- Dynamic association of the Emergency Location Identification Number (ELIN) to the calling phone, for callback purposes. In contrast to the general emergency services scenarios outlined in preceding sections, Cisco Emergency Responder enables the callback to ring the exact phone that initiated the 911 call.
- On-site notification to designated parties (by pager, web page, email, or phone call) to inform them that there is an emergency call in progress. Email, pager, and web page notifications include the calling party name and number, the ERL, and the date and time details associated with the call. Phone notification provides the information about the calling number from which the emergency call was placed.

For more information on ERLs and ELINs, see Emergency Response Location Mapping, page 15-13, and Emergency Location Identification Number Mapping, page 15-14. For more information on Cisco Emergency Responder, see Cisco Emergency Responder Design Considerations, page 15-19, and refer to the Cisco Emergency Responder product documentation available online at

https://www.cisco.com/en/US/products/sw/voicesw/ps842/tsd_products_support_series_home.html

Device Location Discovery Methods in Cisco Emergency Responder

Cisco Emergency Responder uses multiple methods to determine the physical location of a device. Because more specific location discovery results in a shorter time to locate the emergency and administer emergency services, Emergency Responder uses the following methods (listed in priority order) to identify an emergency caller's location:

1. Switch port discovery
2. Access point association
3. IP subnet
4. Static DN assignment
5. Default route

Switch Port Discovery

The primary method for location identification in Cisco Emergency Responder is the detection of an endpoint via Layer 2 discovery at the switch port level. Discovering an endpoint though Layer 2 Cisco Discovery Protocol (CDP) discovery enables Emergency Responder to determine the exact physical location of the calling device based on the physical termination of the network cable to a network jack in a cubicle or office. Although the discovery mechanism of the connected device is reliable, the accuracy of the physical location relies on two main assumptions:

- The wired infrastructure of the enterprise is well established and does not change sporadically, and any wiring closet changes trigger notification to the Emergency Responder administrator indicating what changed.
- The infrastructure is available for Cisco Emergency Responder to browse; that is, Cisco Emergency Responder can establish Simple Network Management Protocol (SNMP) sessions to the underlying network infrastructure and can scan the network ports for the discovery of connected phones.

Once Cisco Emergency Responder discovers the originating port for the call, it associates the call with the pre-established ERL for the location of that port. This process also yields an association with a pre-established ELIN for the location and the selection of the appropriate egress point to the E911 infrastructure, based on the originating ERL.

Access Point Association

Because wireless devices do not have the same discovery capability and tracking characteristics as a wired endpoint, Cisco Emergency Responder tracks wireless clients by using the Location Awareness feature available in Unified CM 11.5 and later releases. The Location Awareness feature allows Emergency Responder to synchronize all deployed access points in Unified CM and to assign the APs to the appropriate ERL. The Location Awareness feature also allows for the updating of mobile device movement between APs.

Emergency Responder is able to track wireless clients across the enterprise through the Location Awareness feature in Unified CM. When a mobile client associates with an AP in the enterprise, the device sends the Basic Service Set Identifier (BSSID) of the AP to Unified CM through call control. Unified CM then updates the database with the new AP association. Periodically, Emergency Responder requests device updates from Unified CM for any device that has updated its AP association since the last request. Emergency Responder receives only the devices that have moved since the last request. In Unified CM 11.5, the request interval is 2 minutes.

IP Subnet

Cisco Emergency Responder also provides the capability to configure ERLs for IP subnets and to assign IP endpoint location by IP address. This capability may be used to locate wireless IP phones, IP softphones, collaboration endpoints that do not support Cisco Discovery Protocol (CDP), and third-party SIP endpoints registered to Cisco Unified CM, which Cisco Emergency Responder cannot locate by connected switch port. It may also be used instead of, or in addition to, connected switch port locations for wired Cisco Collaboration endpoints. If both connected switch port and IP subnet locations are available for a Cisco Collaboration endpoint, Cisco Emergency Responder will prefer the connected switch port location because it is usually more specific than the IP subnet location. Using both connected switch port and IP subnet locations is a best practice because it provides assurance that an appropriate ERL will be assigned, even in case of any delay or error in detecting the connected switch port.

Cisco Emergency Responder allows for the use of two or more ELINs per ERL. The purpose of this enhancement is to cover the specific case of more than one 911 call originating from a given ERL within the same general time period, as illustrated by the following examples.

Example 1

- Phone A and phone B are both located within ERL X, and ERL X is associated with ELIN X.
- Phone A makes a 911 call at 13:00 hours. ELIN X is used to route the call to PSAP X, and PSAP X answers and releases the call. Then, at 13:15 hours, phone B makes a 911 call. ELIN X is again used to route the call to PSAP X.
- PSAP X, after releasing the call from phone B, decides to call back phone A for further details pertaining to phone A's original call. The PSAP dials ELIN X, and gets phone B (instead of the desired phone A).

To work around this situation, Cisco Emergency Responder allows you to define a pool of ELINs for each ERL. This pool provides for the use, in a round-robin fashion, of a distinct ELIN for each successive call. With the definition of two ELINs for ERL X in our example, we now have the situation described in Example 2.

Example 2

- Phone A and phone B are both located within ERL X. ERL X is associated with both ELIN X1 and ELIN X2.
- Phone A makes a 911 call at 13:00 hours. ELIN X1 is used to route the call to PSAP X, and PSAP X answers and releases the call. Then, at 13:15 hours, phone B makes a 911 call, and ELIN X2 is used to route this call to PSAP X.
- PSAP X, after releasing the call from phone B, decides to call back phone A for further details pertaining to phone A's original call. The PSAP dials ELIN X1 and gets phone A.

Of course, if a third 911 call were made but there were only two ELINs for the ERL, the situation would allow for callback functionality to properly reach only the last two callers in the sequence.

High Availability for Emergency Services

It is very important for emergency services to always be available to the user even under the most critical conditions. Therefore, high availability planning must be done carefully when deploying emergency services in an enterprise.

Cisco Emergency Responder supports clustering with a maximum of two servers in active/standby mode. The data is synchronized between the primary and the secondary Cisco Emergency Responder servers. To ensure that calls are routed to the secondary server if the primary server is unavailable, the system administrator must follow certain provisioning guidelines for configuring CTI route points and the directory numbers (DNs) associated to those CTI route points in Cisco Unified CM. For more details on configuration, refer to the *Cisco Emergency Responder Administration Guide*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

If both of the Cisco Emergency Responder servers are unavailable, a local route group (LRG) may be used to route the call to the appropriate PSAP with an appropriate ELIN/ERL (which might be less specific than what Cisco Emergency Responder could have provided). Alternatively, the call may be routed to an internal security office to determine the caller's location. In either case, this provisioning must be done in Cisco Unified CM.

Apart from Cisco Emergency Responder redundancy, Cisco Unified CM redundancy and gateway/trunk redundancy should also be considered to route the 911 emergency calls and to avoid any single point of failure.

Capacity Planning for Cisco Emergency Responder Clustering

In a Cisco Emergency Responder cluster, the quantity of endpoints roaming outside the tracking domain of their home Cisco Emergency Responder group is a scalability factor that must be kept within the limits set forth in the section on *Network Hardware and Software Requirements* in the *Cisco Emergency Responder Administration Guide*, available at:

https://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

For deployments that exceed the Emergency Responder maximum roaming capacity limit (for instance, large campus deployments with multiple Unified CM clusters), phone movement can be tracked by IP subnets. By defining the IP subnets in each of the Cisco Emergency Responder groups and by assigning each ERL with one ELIN per Cisco Emergency Responder group, you can virtually eliminate roaming phones because all phones in the campus will be part of the tracking domain of their respective Cisco Emergency Responder group.

To ensure proper sizing, use the Cisco Collaboration Sizing Tool. This tool is available only to Cisco partners and employees, with appropriate login required, at https://cucst.cloudapps.cisco.com/landing. If you do not have access to this sizing tool, work with your Cisco account team or partner integrator to size your system appropriately.

Design Considerations for 911 Emergency Services

When planning 911 emergency services for multi-line telephone system (MLTS) deployments, first establish all of the physical locations where phone services are needed. The locations can be classified as follows:

- Single building deployments, where all users are located in the same building
- Single campus deployments, where the users are located in a group of buildings situated in close proximity
- Multisite deployments, where users are distributed over a wide geographical area and linked to the call processing site through WAN connectivity

The locations, or type of deployment, affect the criteria used to design and implement 911 services. The following sections describe the key criteria, along with typical and exceptional situations for each. When analyzing and applying these criteria, consider how they are affected by the phone locations in your network.

Emergency Response Location Mapping

The National Emergency Number Association (NENA) has proposed model legislation to be used by state and federal agencies in enacting the rules that govern 911 in enterprise communications systems. One of the concepts in the NENA proposal is that of the emergency response location (ERL), which is defined as:

> *A location to which a 911 emergency response team may be dispatched. The location should be specific enough to provide a reasonable opportunity for the emergency response team to quickly locate a caller anywhere within it.*

Rather than having to identify each endpoint's location individually, the requirement allows for the grouping of endpoints into a "zone," the ERL. The maximum size of the ERL may vary, depending upon local implementation of the legislation, but we will use 7000 square feet (sq ft) as a basis for discussion in this section. (The concepts discussed here are independent of the maximum ERL size that may be allowed in any given state or region.)

An emergency location identification number (ELIN) is associated with each ERL. The ELIN is a fully qualified E.164 number, used to route the call within the E911 network. The ELIN is sent to the E911 network for any 911 call originating from the associated ERL. This process allows more than one phone to be associated with the same fully qualified E.164 number for 911 purposes, and it can be applied to DID and non-DID phones alike.

Note This document does not attempt to present the actual requirements of any legislation. Rather, the information and examples presented here are for the purposes of discussion only. The system planner is responsible for verifying the applicable local requirements.

For example, assume a building has a work area of 70,000 sq ft and 100 endpoints. In planning for 911 functionality, the building can be divided into 10 zones (ERLs) of 7000 sq ft each, and each endpoint can be associated with the ERL where it is located. When a 911 call is made, the ERL (which could be the same for multiple endpoints) is identified by sending the associated ELIN to the PSAP. If the endpoints were evenly distributed in this example, each group of 10 endpoints would have the same ERL and, therefore, the same ELIN.

The various legislations define a minimum number of endpoints (for example, 49) and a minimum work area (for example, 40,000 sq ft) below which the requirements for MLTS 911 are not applicable. But even if the legislation does not require 911 functionality for a given enterprise, it is always best practice to provision for it.

Emergency Location Identification Number Mapping

In general, you must associate a single fully qualified E.164 number, known as the emergency location identification number (ELIN), with each ERL. (However, if using Cisco Emergency Responder, you can configure more than one ELIN per ERL.) The ELIN is used to route the call across the E911 infrastructure and is used by the PSAP as the index into the ALI database.

ELINs must meet the following requirements:

- The ELIN must be routable across the E911 infrastructure. (See the examples in the section on Interface Type, page 15-7.) If an ELIN is not routable, 911 calls from the associated ERL will, at best, be handled according to the default routing programmed in the E911 selective router.
- Once the ERL-to-ELIN mapping of an enterprise is defined, the corresponding ALI records must be established with the LEC so that the ANI and ALI database records serving the PSAP can be updated accurately.
- The ELIN must be reachable from the PSAP for callback purposes.

The ELIN mapping process can be one of the following, depending on the type of interface to the E911 infrastructure for a given ERL:

- Dynamic ANI interface

  With this type of interface, the calling party number identification passed to the network is controlled by the MLTS. The telephony routing table of the MLTS is responsible for associating the correct ELIN with the call, based on the calling endpoint's ERL. In scenarios where Cisco Emergency Responder is not deployed, the calling party number for calls made to 911 can be

modified by Unified CM using transformation masks. For example, all endpoints located in a given ERL can share the same calling search space that lists a partition containing a translation pattern (911) and a calling party transformation mask that would replace the endpoint's CPN with the ELIN for that location. On the other hand, if Cisco Emergency Responder is deployed, calling party number modification should be done on the Emergency Responder system.

- Static ANI interface

  With this type of interface, the calling party number identification passed to the network is controlled by the PSTN. This is the case if the interface is a POTS line. The ELIN is the phone number of the POTS line, and no further manipulation of the phone's calling party identification number is possible.

PSAP Callback

The PSAP might have to reach the caller after completion of the initial conversation or if the caller hangs up before the PSAP operator answers the call. The PSAP's ability to call back relies on the information that it receives with the original incoming call.

The delivery of this information to the PSAP is a two-part process:

1. The Automatic Number Identification (ANI) is first sent to the PSAP. The ANI is the E.164 number used to route the call. In our context, the ANI received at the PSAP is the ELIN that the MLTS sent.
2. The PSAP then uses the ANI to query a database and retrieve the Automatic Location Identification (ALI). The ALI provides the PSAP attendant with information such as:
   - Calling company name
   - Physical address
   - Applicable public safety agency
   - Other optional information, which could include callback information. For example, the phone number of the enterprise's security service could be listed, to aid in the coordination of rescue efforts.

Typical Situation

- The ANI information is used for PSAP callback, which assumes that the ELINs are PSTN dialable numbers.
- The ELINs are PSTN numbers associated with the MLTS. If someone calls the ELIN from the PSTN, the call will terminate on an interface controlled by the MLTS.
- It is the responsibility of the MLTS system administrator to program the call routing so that calls made to any ELIN in the system will ring a phone (or multiple phones) in the immediate vicinity of the associated ERL.
- Once the ERL-to-ELIN mapping is established, it needs be modified only when there are changes to the physical situation of the enterprise. If phones are simply added, moved, or deleted from the system, the ERL-to-ELIN mapping and its associated ANI/ALI database records need not be changed.

Exceptional Situation

- Callback to the immediate vicinity of the originating ERL may be combined with (or even superseded by) routing the callback to an on-site emergency desk, which will assist the PSAP in reaching the original caller and/or provide additional assistance with the emergency situation at hand.
- The situation of the enterprise could change, for example, due to area code splits, city or county service changes requiring a new distribution of the public safety responsibilities, new buildings being added, or any other change that would affect the desired routing of a call for 911 purposes. Any of these events could require changes in the ERL-to-ELIN mapping and the ANI/ALI database records for the enterprise.

Dial Plan Considerations

It is highly desirable to configure a dial plan so that the system easily recognizes emergency calls, irrespective of whether an access code (for example, 9) is used or not. The emergency string for North America is generally 911. Cisco strongly recommends that you configure the system to recognize both the strings 911 and 9911.

Cisco also strongly recommends that you explicitly mark the emergency route patterns with Urgent Priority so that Unified CM does *not* wait for the inter-digit timeout (Timer T.302) before routing the call.

Other emergency call strings may be supported concurrently on your system. Cisco highly recommends that you provide your system users with training on the selected emergency call strings.

Also, it is highly desirable that users be trained to react appropriately if they dial the emergency string by mistake. In North America, 911 may be dialed in error by users trying to access a long distance number through the use of 9 as an access code. In such a case, the user should remain on the line to confirm that there is no emergency, and therefore no need to dispatch emergency personnel. Cisco Emergency Responder's on-site notification capabilities can help in identifying the phone at the origin of such spurious 911 calls by providing detailed accounts of all calls made to 911, including calls made by mistake. If the emergency dispatch center cannot confirm that a call to 911 was accidental, then emergency services must be dispatched to the calling location. More than three emergency services dispatches to a single customer in a month often times will result in a fine to the company.

In a multisite deployment, the dial plan configuration should ensure that the emergency calls are always routed through the PSTN gateway local to the site, thereby making sure that the emergency call is routed to the nearest PSAP within the jurisdiction. One of the mechanism to achieve this could be to use the Local Route Group feature of Cisco Unified CM. In the case of multisite deployments with centralized PSTN access, local call routing to the PSAP is not possible. For deployments with centralized PSTN access, the Unified CM administrator must verify that the PSTN provider will route emergency calls to the proper PSAP based on ANI or ELIN. If the service provider cannot provide emergency call routing services for multiple sites, then any site not included in E911 coverage must have a location connection (an analog line) or the centralized PSTN access must support 911 call delivery for remote sites (a SIP trunk). (See the examples in the section on Interface Type, page 15-7.)

Also, in a multisite deployment it is very important to make sure that the emergency number is always reachable and routed through the local PSTN gateway for the mobility users (extension mobility and device mobility) independent of the implemented Class of Service (CoS). If the site/device approach is being used, the device calling search space (CSS) could be used to route the emergency calls.

Cisco recommends enabling Calling Party Modification on Cisco Emergency Responder. When this feature is enabled, the calling party number is replaced with the ELIN by Cisco Emergency Responder for the emergency call. If Calling Party Modification is not enabled, either the DID will be sent to the PSAP or Cisco Unified CM must be configured to replace the calling party with the ELIN defined on the route pattern or the gateway.

Gateway Considerations

Consider the following factors when selecting the gateways to handle emergency calls for your system:

- Gateway Placement, page 15-17
- Gateway Blocking, page 15-17
- Answer Supervision, page 15-18
- Answer Supervision, page 15-18

Gateway Placement

Within the local exchange carrier (LEC) networks, 911 calls are routed over a locally significant infrastructure based on the origin of the call. The serving Class 5 switches are connected either directly to the relevant PSAP for their location or to an E911 selective router, which itself is connected to a group of PSAPs significant for its region.

With Cisco's IP-based enterprise communications architecture, it is possible to route calls on-net to gateways that are remotely situated. As an example, an endpoint located in San Francisco could have its calls carried over an IP network to a gateway situated in San Jose, and then sent to the LEC's network.

For 911 calls, it is critical to choose the egress point to the LEC network so that emergency calls are routed to the appropriate local PSAP. In the example above, a 911 call from the San Francisco endpoint, if routed to a San Jose gateway, could not reach the San Francisco PSAP because the San Jose LEC switch receiving the call does not have a link to the E911 selective router serving the San Francisco PSAP. Furthermore, the San Jose area 911 infrastructure would not be able to route the call based on a San Francisco calling party number.

As a general rule, route 911 calls to a gateway physically located with the originating endpoint. Contact the LEC to explore the possibility of using a common gateway to aggregate the 911 calls from multiple locations. Be aware that, even if the 911 network in a given region lends itself to using a centralized gateway for 911 calls, it might be preferable to rely on gateways located with the calling phones to prevent 911 call routing from being impacted during WAN failures.

Gateway Blocking

It is highly desirable to protect 911 calls from "all trunks busy" situations. If a 911 call needs to be connected, it should be allowed to proceed even if other types of calls are blocked due to lack of trunking resources. To provide for such situations, you can dedicate an explicit trunk group just for 911 calls.

It is acceptable to route emergency calls exclusively to an emergency trunk group. Another approach is to send emergency calls to the same trunk group as the regular PSTN calls (if the interface permits it), with an alternative path to a dedicated emergency trunk group. The latter approach allows for the most flexibility.

As an example, we can point emergency calls to a PRI trunk group, with an alternate path (reserved exclusively for emergency calls) to POTS lines for overflow conditions. If we put 2 POTS lines in the alternate trunk group, we are guaranteeing that a minimum of two simultaneous 911 calls can be routed, in addition to any calls that were allowed in the main trunk group.

If the preferred gateway becomes unavailable, it may be acceptable to overflow emergency calls to an alternate number so that an alternate gateway is used. For example, in North America calls dialed as 911 could overflow to an E.164 (non-911) local emergency number. This approach does not take advantage of the North American 911 network infrastructure (that is, there is no selective routing, ANI, or ALI services), and it should be used only if it is acceptable to the applicable public safety authorities and only as a last resort to avoid rejecting the emergency call due to a lack of network resources.

Answer Supervision

Under normal conditions, calls made to an emergency number should return answer supervision upon connection to the PSAP. The answer supervision may, as with any other call, trigger the full-duplex audio connection between the on-net caller and the egress interface to the LEC's network.

With some North American LECs, answer supervision might not be returned when a "free" call is placed. This may be the case for some toll-free numbers (for example, 800 numbers). In exceptional situations, because emergency calls are considered "free" calls, answer supervision might not be returned upon connection to the PSAP. You can detect this situation simply by making a 911 test call. Upon connection to the PSAP, if audio is present, the call timer should record the duration of the ongoing call; if the call timer is absent, it is very likely that answer supervision was not returned. If answer supervision is not returned, Cisco highly recommends that you contact the LEC and report this situation because it is most likely not the desired functionality.

If this situation cannot be rectified by the Local Exchange Carrier, it would be advisable to configure the egress gateway *not* to require answer supervision when calls are placed to the LEC's network, and to cut through the audio in both directions so that progress indicator tones, intercept messages, and communications with the PSAP are possible even if answer supervision is not returned.

By default, Cisco IOS-based H.323 gateways must receive answer supervision in order to connect audio in both directions. To forego the need for answer supervision on these gateways, use the following commands:

- **progress_ind alert enable 8**

  This command provides the equivalent of receiving a progress indicator value of 8 (in-band information now available) when alerting is received. This command allows the POTS side of the gateway to connect audio toward the origin of the call.

- **voice rtp send-recv**

  This command allows audio cut-through in both backward and forward directions before a connect message is received from the destination switch. This command affects all Voice over IP (VoIP) calls when it is enabled.

Be advised that, in situations where answer supervision is not provided, the call detail records (CDRs) will not accurately reflect the connect time or duration of 911 calls. This inaccuracy can impede the ability of a call reporting system to document the relevant statistics properly for 911 calls.

In all cases, Cisco highly recommends that you test 911 call functionality from all call paths and verify that answer supervision is returned upon connection to the PSAP.

Cisco Emergency Responder Design Considerations

Device mobility brings about special design considerations for emergency calls. Cisco Emergency Responder (Emergency Responder) can be used to track device mobility and to adapt the system's routing of emergency calls based on a device's dynamic physical location.

Device Mobility Across Call Admission Control Locations

In a centralized call processing deployment, Cisco Emergency Responder can detect Cisco endpoint relocation and reassign relocated endpoints to appropriate ERLs automatically. However, Cisco Unified CM location-based call admission control for a relocated endpoint might not properly account for the WAN bandwidth usage of the phone in the new location, yielding possible over-subscription or under-subscription of WAN bandwidth resources. For example, if you physically move a phone from Branch A to Branch B, the endpoint's call admission control location remains the same (Location_A), and it is possible that calls made to 911 from that endpoint would be blocked due to call admission control denial if all available bandwidth to Location_A is in use for other calls. To avoid such blocking of calls, manual intervention might be required to adapt the device's location and region parameters.

Cisco Unified CM device mobility provides a way to update the endpoint's configuration automatically (including its calling search space and location information) in Unified CM to reflect its new physical location. If device mobility is not used, manual configuration changes may be necessary in Cisco Unified CM.

For more details on the Device Mobility feature, refer to the section on Device Mobility, page 21-14.

Default Emergency Response Location

If Cisco Emergency Responder cannot directly determine the physical location of an endpoint, it assigns a default emergency response location (ERL) to the call. The default ERL points all such calls to a specific PSAP. Although there is no universal recommendation as to where calls should be sent when this situation occurs, it is usually desirable to choose a PSAP that is centrally located and that offers the largest public safety jurisdiction. It is also advisable to populate the ALI records of the default ERL's emergency location identification numbers (ELINs) with contact information for the enterprise's emergency numbers and to offer information about the uncertainty of the caller's location. In addition, it is advisable to mark those ALI records with a note that a default routing of the emergency call has occurred. Alternatively, the call may be routed to an internal security office to determine the caller's location.

Cisco Emergency Responder and Location Awareness for Wireless Clients

Cisco Emergency Responder 11.5 and later releases can track wireless endpoints and clients to an access point in the enterprise. To minimize configuration changes in Cisco Emergency Responder, all access points must be synchronized from Cisco Unified Communications Manager. The synchronization process also handles any access point additions, updates, or removals that occur in Cisco Unified CM. Any access point changes in Unified CM are seen in Emergency Responder within 2 minutes of the change. Access points cannot be defined within Emergency Responder, and all access points that are to be used for location identification in Emergency Responder must be defined in Unified CM. For access point management, Unified CM uses the Cisco Wireless LAN Controller Synchronization Service to automatically synchronize access points into the Unified CM database. The Cisco Wireless LAN Controller Synchronization Service integrates with Cisco Wireless LAN Controllers (WLCs) for access point information. If another vendor is used for WLC services, then the access points must be bulk imported into the Cisco Unified CM database using the Bulk Administration Tool (BAT).

For a mobile client or wireless device to be associated to a wireless access point, the client or device must send the Basic Service Set Identifier (BSSID) of the associated access point to Cisco Unified CM. Due to the frequency of updates that a mobile client can generate, Unified CM limits the rate of location updates from mobile devices and wireless clients to 90 updates per second per node. If location updates exceed this rate for a sustained period of time, Unified CM defers further updates with a 480 "Busy Here" message. The client responds by waiting a period of time before sending the location update again. The amount of delay before sending the update again dependents on the client and not on Cisco Emergency Responder or Unified CM.

When a mobile client or wireless device updates its location in Cisco Unified CM, the update is reflected in Cisco Emergency Responder in less than 2 minutes.

Cisco Emergency Responder and Extension Mobility

Cisco Emergency Responder supports Extension Mobility within a Cisco Unified CM cluster. It can also support Extension Mobility Cross-Cluster (EMCC), provided that both Cisco Unified CM clusters are supported either by a common Cisco Emergency Responder server or group, or by two Cisco Emergency Responder servers or groups configured as a Cisco Emergency Responder cluster. In either case, the Cisco Unified CM clusters must not be configured to use the Adjunct Calling Search Space (CSS) associated with EMCC for 911 calls, but must be configured to use Cisco Emergency Responder for all 911 calls in both Cisco Unified CM clusters.

Cisco Emergency Responder and Video

Cisco Emergency Responder can discover Cisco Video Collaboration endpoints in the following ways, depending on their capabilities:

- Video Collaboration Endpoints that Support CDP, page 15-20
- Video Collaboration Endpoints that Do Not Support CDP, page 15-21

Regardless of which way the video endpoints are discovered, it is important to note that video is not supported as media for emergency calling to the PSAP.

Note The topics discussed in this chapter apply to Cisco Emergency Responder only when it is used in conjunction with Cisco Unified Communications Manager (Unified CM). Cisco TelePresence Video Communication Server (VCS) currently does not support emergency services.

Video Collaboration Endpoints that Support CDP

For video collaboration endpoints that support Cisco Discovery Protocol (CDP) and that are within the corporate premises, Cisco recommends treating them like any other collaboration endpoints tracked by Cisco Emergency Responder through CDP, as described by the Emergency Responder switch configuration information in the latest version of the *Cisco Emergency Responder Administration Guide,* available at

https://www.cisco.com/en/US/partner/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

For video collaboration endpoints with CDP support that are outside the corporate premises, Cisco recommends treating them like voice collaboration endpoints as described in the information for off-premises support of IP phones in the latest version of the *Off-Premise Location Management User Guide for Cisco Emergency Responder*, available at

https://www.cisco.com/en/US/partner/products/sw/voicesw/ps842/products_user_guide_list.html

Video Collaboration Endpoints that Do Not Support CDP

For video collaboration endpoints that do not support Cisco Discovery Protocol (CDP), Cisco recommends using a dedicated line for a voice collaboration endpoint. If you require tracking of the video collaboration endpoint, then Cisco recommends configuring an IP subnet ERL as described in the information about setting up IP subnet-based ERLs found in the latest version of the *Cisco Emergency Responder Administration Guide,* available at

https://www.cisco.com/en/US/partner/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

Cisco Emergency Responder and Off-Premises Endpoints

In cases where endpoints are located outside of the enterprise boundary but connect back to the enterprise using VPN or VPN-less solutions (for example, Cisco Expressway mobile and remote access from a home office or hotel), Cisco Emergency Responder will not be able to determine the location of the caller. Furthermore, it is unlikely that the system would have a gateway properly situated to allow sending the call to the appropriate PSAP for the caller's location.

It is a matter of enterprise policy to allow or not to allow the use of off-premises endpoints for 911 calls through the enterprise. It might be advisable to disallow 911 calls by policy for those endpoints that connect over the Internet through VPN or Cisco Expressway. Nevertheless, if such a user were to call 911, the best-effort system response would be to route the call to either an on-site security force or a large PSAP close to the system's main site.

The following paragraph is an example notice that you could issue to users to warn them that emergency call functionality is not guaranteed for off-premises endpoints and users:

> *Emergency calls should be placed from devices that are located at the site for which they are configured (for example, your office). A local safety authority might not answer an emergency call placed from a device that has been removed from its configured site. If you must use this device for emergency calls while away from your configured site, be prepared to provide the answering public safety authority with specific information regarding your location. Use a device that is locally configured to the site (for example, your hotel phone or your home phone) for emergency calls when traveling or telecommuting.*

Cisco Emergency Responder also supports integration with Intrado V9-1-1, an emergency call delivery service that can reach almost any PSAP in the United States. With the combination of Cisco Emergency Responder and Intrado V9-1-1, users of IP phones and softphones outside the enterprise can update their locations by using the display screen on most Cisco IP Phones and Cisco IP Communicator or by using a web page provided by Cisco Emergency Responder. Emergency calls from an off-premises location will then be delivered through Cisco Emergency Responder to Intrado and then to the appropriate PSAP for the caller's location.

Test Calls

For any enterprise telephony system, it is a good idea to test 911 call functionality, not only after the initial installation, but regularly, as a preventive measure.

The following suggestions can help you carry out the testing:

- Contact the PSAP to ask for permission before doing any tests, and provide them with the contact information of the individuals making the tests.
- During each call, indicate that it is *not* an actual emergency, just a test.
- Confirm the ANI and ALI that the call taker has on their screen.

- Confirm the PSAP to which the call was routed.
- Confirm that answer supervision was received by looking at the call duration timer on the endpoint. An active call timer is an indication that answer supervision is working properly.

PSAP Callback to Shared Directory Numbers

Cisco Emergency Responder handles the routing of inbound calls made to emergency location identification numbers (ELINs). In cases where the line from which a 911 call was made is a shared directory number, the PSAP callback will cause all shared directory number appearances to ring. Any of the shared appearances can then answer the call, which means that it may not be the phone from which the 911 call originated.

In Cisco Unified CM 11.5 and later releases, a PSAP callback to a shared DN will ring only the device that placed the call to the PSAP. Unified CM will override device and line settings (such as Call Forward All and Do Not Disturb) to deliver the callback from emergency services.

Cisco Emergency Responder Deployment Models

Enterprise communications systems based on multiple Unified CM clusters can benefit from the functionality of Cisco Emergency Responder (Emergency Responder).

The *Cisco Emergency Responder Administration Guide* provides detailed descriptions of the terms used herein, as well as the background information required to support the following discussion. Of specific interest is the chapter on *Planning for Cisco Emergency Responder*. This documentation is available at

https://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

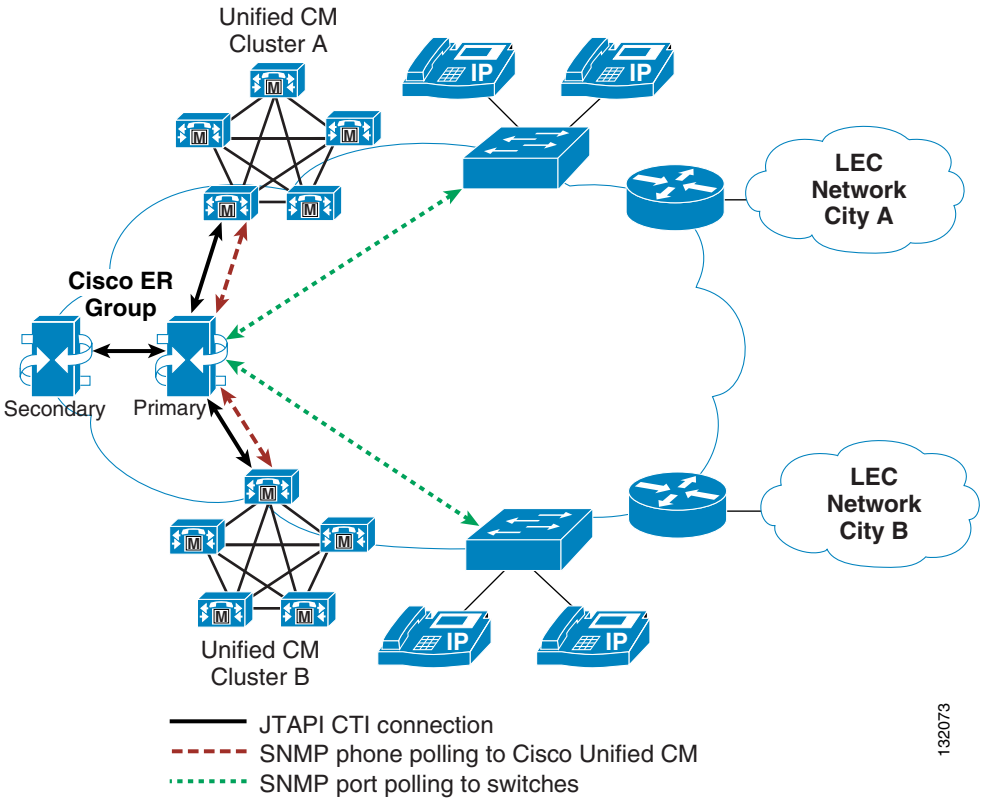
Note Cisco Emergency Responder does not support Cisco Unified Communications Manager Express (Unified CME) or Survivable Remote Site Telephony (SRST). In case of SRST deployment, configure the appropriate dial-peer to route the 911 calls to the PSTN with the published site number. Unified CME natively supports E911.

Single Cisco Emergency Responder Group

A single Emergency Responder group can be deployed to handle emergency calls from two or more Unified CM clusters. The design goal is to ensure that an emergency call from any phone is routed to the Cisco Emergency Responder group, which will assign an ELIN and route the call to the appropriate gateway based on the endpoint's location.

One advantage of using a single Cisco Emergency Responder group is that all ERLs and ELINs are configured into a single system. An endpoint registered on any cluster will be located by the single Cisco Emergency Responder group because that group is responsible for polling all of the system's access switches. Figure 15-3 illustrates a single Cisco Emergency Responder group interfaced with two Unified CM clusters.

Figure 15-3 A Single Cisco Emergency Responder Group Connected to Two Unified CM Clusters



The single Cisco Emergency Responder group in Figure 15-3 interfaces with the following components:

- Each Unified CM cluster, via SNMP, to collect information about their respective configured endpoints.
- Enterprise access switches, via SNMP, where IP telephony endpoints are connected. This connection is not required if the endpoint locations are being identified based on IP subnets. For details on configuring IP subnet-based ERLs, refer to the *Cisco Emergency Responder Configuration* chapter in the *Cisco Emergency Responder Administration Guide*, available at

  https://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html
- Each Unified CM cluster, via JTAPI, to allow for the call processing required by any endpoint that dials 911 – for example, identification of the calling endpoint's ERL, assignment of the ELIN, redirection of the call to the proper gateway (based on the calling endpoint's location), and the handling of the PSAP callback functionality.
- Each Unified CM cluster, via SNMP, to collect access point information from a Cisco Wireless LAN Controller (WLC).

The version of the JTAPI interface used by Cisco Emergency Responder is determined by the version of the Unified CM software to which it is connected. At system initialization, Cisco Emergency Responder interrogates the Unified CM cluster and loads the appropriate JTAPI Telephony Service Provider (TSP). Because there can be only one version of JTAPI TSP on the Cisco Emergency Responder server, all Unified CM clusters to which a single Cisco Emergency Responder group is interfaced *must* run the same version of Unified CM software.

For some deployments, this software version requirement might present some difficulties. For instance, during a Unified CM upgrade, different clusters will be running different versions of software, and some of the clusters will be running a version of JTAPI that is not compatible with the version running on the Cisco Emergency Responder servers. When this situation occurs, emergency calls from the cluster running a version of JTAPI different than that of the Cisco Emergency Responder group might receive the call treatment provided by the call forward settings of the emergency number's CTI Route Point.

When considering if a single Cisco Emergency Responder group is appropriate for multiple Unified CM clusters, apply the following guidelines:

- Make Unified CM upgrades during an acceptable maintenance window when emergency call volumes are as low as possible (for example, after hours, when system use is at a minimum).
- Use a single Cisco Emergency Responder group only if the quantity and size of the clusters allow for minimizing the amount of time when dissimilar versions of JTAPI are in use during software upgrades.

For example, a deployment with one large eight-server cluster in parallel with a small two-server cluster could be considered for use with a single Cisco Emergency Responder group. In this case, it would be best to upgrade the large cluster first, thus minimizing the number of users (those served by the small cluster) that might be without Cisco Emergency Responder service during the maintenance window of the upgrade. Furthermore, the small cluster's users can more appropriately be served by the temporary static routing of emergency calls in effect while Cisco Emergency Responder is not reachable because they can be identified by the single ERL/ELIN assigned to all non-ER calls made during that time.

Multiple Cisco Emergency Responder Groups

Multiple Cisco Emergency Responder groups can also be deployed to support multi-cluster systems. In this case, each ER group interfaces with the following components:

- A Unified CM cluster via the following methods:
  - SNMP, to collect information about its configured endpoints
  - JTAPI, to allow for the call processing associated with redirection of the call to the proper gateway or, in the case of roaming endpoints, the proper Unified CM cluster
- The access switches (via SNMP) to which most of the endpoints associated with the Unified CM of the Cisco Emergency Responder group are most likely to be connected
- Each Unified CM cluster (via SNMP) to collect Access Point information from a Cisco Wireless LAN Controller (WLC)

This approach allows Unified CM clusters to run different versions of software because each is interfaced to a separate Cisco Emergency Responder group.

To allow endpoints to roam between various parts of the network and still be tracked by Cisco Emergency Responder, you might have to configure the Cisco Emergency Responder groups into a Cisco Emergency Responder cluster. For details on Cisco Emergency Responder clusters and groups, refer to the chapter on *Planning for Cisco Emergency Responder* in the *Cisco Emergency Responder Administration Guide*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

Figure 15-4 presents a sample topology illustrating some of the basic concepts behind Cisco Emergency Responder clustering.

Figure 15-4 Multiple Cisco Emergency Responder Groups

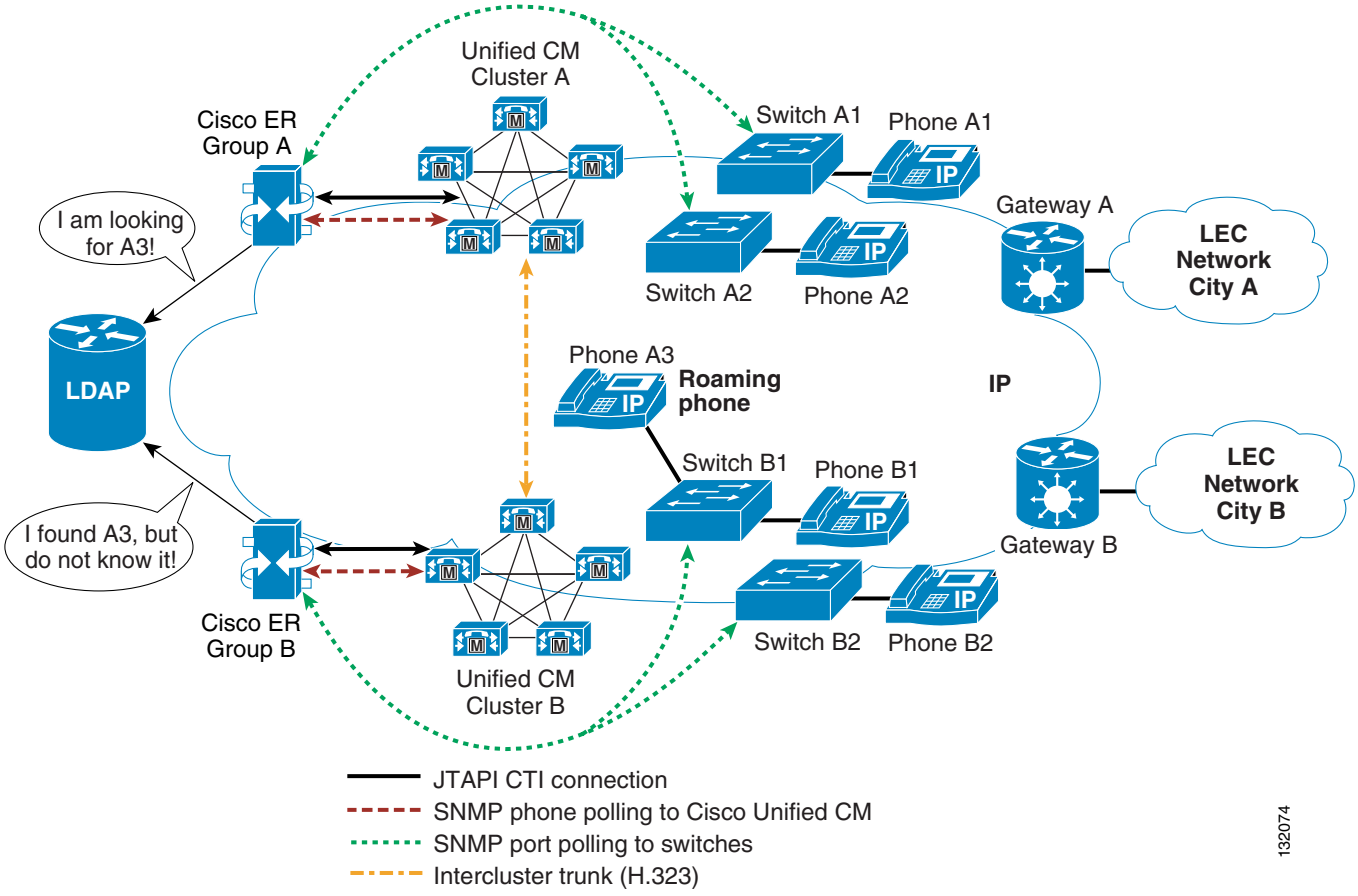


Figure 15-4 illustrates the following topology:

- Cisco Emergency Responder group A is interfaced to Unified CM cluster A to access switches A1 and A2, and it is deemed to be the home Cisco Emergency Responder group of all endpoints registered to Unified CM cluster A.
- Likewise, Cisco Emergency Responder group B is interfaced to Unified CM cluster B to access switches B1 and B2, and it is deemed to be the home Cisco Emergency Responder group of all endpoints registered to Unified CM cluster B.

Endpoint Movements Within the Tracking Domain of a Cisco Emergency Responder Group

The emergency call processing for endpoints moving between access switches controlled by the same home Cisco Emergency Responder group is the same as the processing done for a deployment with a single Unified CM cluster. For example, an endpoint moving between access switches A1 and A2 remains registered with Unified CM cluster A, and its location is determined by Cisco Emergency Responder group A both before and after the move. The endpoint is still under full control of Cisco Emergency Responder group A, for both the discovery of the endpoint by Unified CM cluster A and the determination of the endpoint's location on switch A2 by Cisco Emergency Responder. The endpoint is therefore not considered to be an unlocated phone.

Endpoint Movements Between the Various Tracking Domains of a Cisco Emergency Responder Cluster

A Cisco Emergency Responder cluster is essentially a collection of Cisco Emergency Responder groups that share location information. Each group shares the location of any endpoint it finds on an access switch or in an IP subnet.

Cisco Emergency Responder groups also share information about endpoints that cannot be located within a Cisco Emergency Responder group's tracking domain (in switches or IP subnets) but which are known to be registered in the group's associated Unified CM cluster. Such endpoints are deemed *unlocated*.

If an endpoint is roaming between access switches monitored by different Cisco Emergency Responder groups, those groups must be configured in a Cisco Emergency Responder cluster so they can exchange information about the endpoint's location. For example, endpoint A3 is registered with Unified CM cluster A, but it is connected to an access switch controlled by Cisco Emergency Responder group B. Cisco Emergency Responder group A is aware that endpoint A3 is registered with Unified CM cluster A, but group A cannot locate endpoint A3 in any of the site A switches. Therefore, endpoint A3 is deemed *unlocated* by Cisco Emergency Responder group A.

Cisco Emergency Responder group B, on the other hand, has detected the presence of endpoint A3 in one of the switches that it monitors. Because the endpoint is not registered with Unified CM cluster B, endpoint A3 is advertised through the Cisco Emergency Responder database as an *unknown* endpoint.

Because the two Cisco Emergency Responder groups are communicating through a replicated database table, they can determine that Cisco Emergency Responder group B's *unknown* endpoint A3 is the same as Cisco Emergency Responder group A's *unlocated* endpoint A3.

The Unlocated Phone page in Cisco Emergency Responder group A will display the endpoint's MAC address along with the remote Cisco Emergency Responder group (in this, case Cisco Emergency Responder group B).

Emergency Call Routing within a Cisco Emergency Responder Cluster

Cisco Emergency Responder clustering also relies on route patterns that allow emergency calls to be redirected between pairs consisting of a Unified CM cluster and a Cisco Emergency Responder. For more details, refer to the section on *Creating Route Patterns for Inter-Cisco Emergency Responder Group Communications* in the *Cisco Emergency Responder Administration Guide*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

If endpoint A3 places an emergency call, the call signaling flow will be as follows:

1. Endpoint A3 sends the emergency call string to Unified CM cluster A for processing.
2. Unified CM cluster A sends the call to Cisco Emergency Responder group A for redirection.
3. Cisco Emergency Responder group A determines that endpoint A3 is located in Cisco Emergency Responder group B's tracking domain, so it redirects the call to a route pattern that points to Unified CM cluster B.
4. Unified CM cluster A sends the call to Unified CM cluster B over a SIP trunk or an intercluster trunk.
5. Unified CM cluster B sends the call to Cisco Emergency Responder group B for redirection.
6. Cisco Emergency Responder group B identifies the ERL and ELIN associated with endpoint A3's location (based on calling party number) and redirects the call to Unified CM cluster B. The calling number is transformed into the ELIN associated with the ERL of endpoint A3, and the called number is modified to route the call to the proper gateway.

7. Unified CM cluster B routes the call according to the new called number information obtained from Cisco Emergency Responder group B.
8. Unified CM cluster B sends the call out the gateway toward the Emergency PSTN network.

Note The ERL and ELIN match in step 6 is based on the calling party number of the endpoint placing the call to 911. If the SIP trunk or intercluster trunk modifies the calling party number (perhaps to a full +E.164 number), then the Emergency Responder for Group B will not be able to match the calling party number over the trunk with the directory number learned from the access switch Cisco Discovery Protocol (CDP) neighbor. Therefore, emergency calls that traverse a SIP trunk or intercluster trunk must not undergo any calling party transformations.

WAN Deployment of Cisco Emergency Responder

Cisco Emergency Responder supports two main sites using clustering over the WAN. Install one Emergency Responder server in each site, and configure one server as the publisher and the other server as a subscriber. The Emergency Responder publisher should be located with the primary Unified CM CTI Manager, and the Emergency Responder subscriber should be located with the secondary Unified CM CTI Manager. Any Unified CM server remote from either Emergency Responder server must be within 80 ms round-trip time (RTT) of both Emergency Responder servers. The Emergency Responder publisher and subscriber must also be within 80 ms RTT of each other. The minimum bandwidth required between the Cisco Emergency Responder servers is 1.544 Mbps.

Emergency Call Routing Using Unified CM Native Emergency Call Routing

Customers that require accurate location identification but have a single site or small number of locations that need to be identified, can use the Cisco Unified Communications Manager Native Emergency Call Routing feature. The Native Emergency Call Routing feature allows an administrator to define Emergency Location Identification Numbers (ELINs) at the device pool level or device level so that a device's location can be determined and identified at the public safety answering point (PSAP).

Cisco Unified CM Native Emergency Call Routing provides the following functionality:

- ELIN association based on a static device assignment or device pool assignment
- Dynamic association of the ELIN to the calling phone for callback purposes
- For mobile devices, Device Mobility Groups used to track mobile devices with Native Emergency Call Routing
- Automatic replacement of the calling party number with the appropriate ELIN
- Routing emergency calls to the appropriate gateway for emergency call completion

Design Considerations for 911 Native Emergency Call Routing Services

When designing an emergency call routing plan using Cisco Unified CM Native Emergency Call Routing services, give special consideration to the boundaries of an emergency location inside a building. An emergency location should be an identifiable location with physical or logical boundaries to reduce the amount of time for emergency services to locate an individual in an emergency situation. Examples of physical or logical boundaries can include: a single floor of a building, a lab, an office, or a directional floor indicator (for example, West side of first floor).

The design for Native Emergency Call Routing requires an ELIN to be defined and assigned to devices or devices pools, but the Native Emergency Call Routing feature does not allow the administrator to define the ERL information to be associated with the ELIN. The ERL definition for a given ELIN must be done outside of Cisco Unified CM and uploaded to the local PSAP per the instructions provided by the local exchange carrier when establishing E911 services.

Similar to a Cisco Emergency Responder deployment, Native Emergency Call Routing can support multiple unique and concurrent calls to emergency services from the same location. Native Emergency Call Routing allows the creation of a pool of ELINs that are associated with an emergency location. The number of locations that can be defined is based on the number of ELINs assigned to an individual Emergency Location (ELIN) Group. Native Emergency Call Routing supports a maximum of 100 ELINs. If the deployment requires only one concurrent call per a location, then the system can support 100 unique Emergency Location Groups. If the deployment requires the ability to track 2 concurrent callers from the same location, then the administrator must define 2 ELINs for a single Emergency Location (ELIN) Group. If 2 ELINs are required for a single location, Unified CM will be able to support 50 locations (2 ELINs * 50 ERLs = 100 ELINs). Using more ELINs to support concurrent and uniquely identified callers from a location will reduce the total number of locations that can be defined. The following formula can be used to determine the maximum number of locations that can be defined based on the number of concurrent and unique callers from an ERL:

100/(Number of unique and concurrent callers per ERL) = Max ERLs

ELINs are not required to be the same for each Emergency Location (ELIN) Group. If one ERL covers a high-density user population, the Emergency Location (ELIN) Group may contain 4 ELINs to support 4 concurrent and unique emergency callers. But if the same building has a large lab floor or warehouse that has a small number of regular employees, then that location might have only one ELIN assigned to the Emergency Location (ELIN) Group.

If the PSAP needs to call back and get additional information from the caller, the call will return to Unified CM using the ELIN that originated the call. To route the return call correctly, the dial plan must be configured so that the inbound called number matches the ELIN defined in Unified CM. If the inbound trunk delivers only the last 5 digits of the called party, then the administrator must include a translation pattern to expand the collected digits to match the ELIN. For proper return call operation, the called number must match exactly the ELIN number as defined in Unified CM. Although ELINs can be any number in a customer's DID range, Cisco recommends keeping the ELIN numbers contiguous to use as few call translation patterns as possible.

ALI Formats

In multi-cluster configurations, there might be instances where the physical locations of ERLs and ELINs defined in a single Cisco Emergency Responder group span the territory of more than one phone company. This condition can lead to situations where records destined for different phone companies have to be extracted from a common file that contains records for multiple LECs.

Cisco Emergency Responder exports this information in ALI records that conform to National Emergency Number Association (NENA) 2.0, 2.1, and 3.0 formats. However, many service providers do not use NENA standards. In such cases, you can use the ALI Formatting Tool (AFT) to modify the ALI records generated by Cisco Emergency Responder so that they conform to the formats specified by the service provider. The service provider can then use the reformatted file to update their ALI database.

The ALI Formatting Tool (AFT) enables you to perform the following functions:

- Select a record and update the values of the ALI fields. AFT allows you to edit the ALI fields to customize them to meet the requirements of various service providers. The service provider can then read the reformatted ALI files and use them to update their ELIN records.
- Perform bulk updates on multiple ALI records. Using the bulk update feature, you can apply common changes to all the records that you have selected.
- Selectively export ALI records based on area code, city code, or a four-digit directory number. By selecting to export all the ALI records in an area code, for example, you can quickly access all the ELIN records for each service provider, thereby easily supporting multiple service providers.

Given the flexibility of the AFT, a single Cisco Emergency Responder group can export ALI records in multiple ALI database formats. For a Cisco Emergency Responder group serving a Unified CM cluster with sites in the territories of two LECs, the basic approach is as follows:

1. Obtain an ALI record file output from Cisco Emergency Responder in standard NENA format. This file contains the records destined for multiple LECs.
2. Make a copy of the original file for each required ALI format (one copy per LEC).
3. Using the AFT of the first LEC (for example, LEC-A), load a copy of the NENA-formatted file and delete the records of all the ELINs associated with the other LECs. The information to delete can usually be identified by NPA (or area code).
4. Save the resulting file in the required ALI format for LEC-A, and name the file accordingly.
5. Repeat steps 3 and 4 for each LEC.

For more information about the ALI formatting tools, refer to the online documentation available at

https://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

For LECs not listed at this URL, the output from Emergency Responder can be formatted using standard text file editing tools, such as spreadsheet programs and standard text editors.

Directory Integration and Identity Management

Revised: March 1, 2018

Identity management is a fundamental concept required in any application. Identity management involves the management of individual principals and the authentication and authorization of these principals. Traditionally each application handled identity management individually. This led to the situation that users had to authenticate against every individual application. Centralizing identity management, authentication, and authorization helps greatly to improve the user experience by providing services such as single sign-on (SSO).

The first step of centralizing identity management is to centralize storage of information about principals in an enterprise. These centralized enterprise-wide datastores are commonly known as directories.

Directories are specialized databases that are optimized for a high number of reads and searches, and occasional writes and updates. Directories typically store data that does not change often, such as employee information, user privileges, and group membership on the corporate network.

Directories are extensible, meaning that the type of information stored can be modified and extended. The term *directory schema* defines the type of information stored, its container (or attribute), and its relationship to users and resources.

The Lightweight Directory Access Protocol (LDAP) provides applications with a standard method for accessing and potentially modifying the information stored in the directory. This capability enables companies to centralize all user information in a single repository available to several applications, with a remarkable reduction in maintenance costs through the ease of adds, moves, and changes.

This chapter covers the main design principles for integrating a Cisco Unified Communications system based on Cisco Unified Communications Manager (Unified CM) with a corporate LDAP directory. The main topics include:

- What is Directory Integration?, page 16-3

  This section analyzes the various requirements for integration with a corporate LDAP directory in a typical enterprise IT organization.

- Directory Access for Unified Communications Endpoints, page 16-4

  This section describes the technical solution to enable directory access for Cisco Unified Communications endpoints and provides design best-practices around it.

- Directory Integration with Unified CM, page 16-7

  This section describes the technical solutions and provides design considerations for directory integration with Cisco Unified CM, including the LDAP synchronization and LDAP authentication functions.

- Directory Integration for VCS Registered Endpoints, page 16-33

  This section briefly introduces the technical solution to enable directory access for video endpoints registered to the Cisco TelePresence Video Communication Server (VCS).

- Identity Management Architecture Overview, page 16-33

  This section describes the identity management architecture.

- Single Sign-On (SSO), page 16-35

  This section provides an overview of SAML 2.0 single sign-on (SSO).

- Authorization Framework, page 16-45

  This section describes the OAuth authorization service available in Cisco Unified CM.

The considerations presented in this chapter apply to Cisco Unified CM as well as the following applications bundled with it: Cisco Extension Mobility, Cisco Unified Communications Manager Assistant, WebDialer, Bulk Administration Tool, and Real-Time Monitoring Tool.

For Cisco Unity, refer to the *Cisco Unity Design Guide* and to the following white papers: *Cisco Unity Data and the Directory*, *Active Directory Capacity Planning*, and *Cisco Unity Data Architecture and How Cisco Unity Works*, also available at

https://www.cisco.com

What's New in This Chapter

Table 16-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

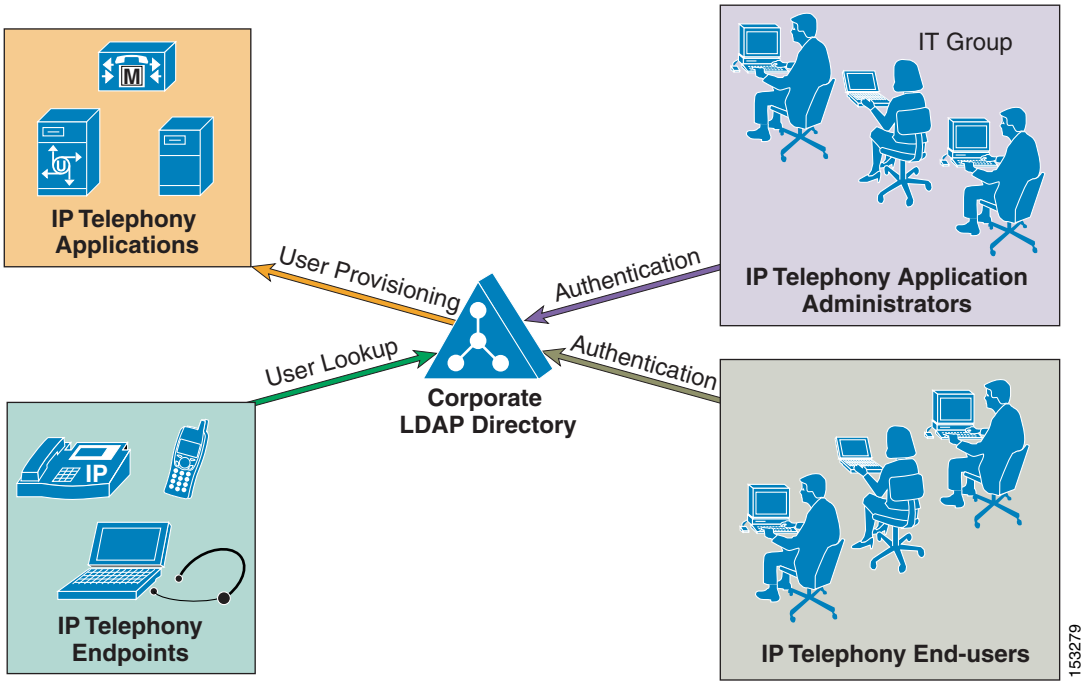
Table 16-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Directory access using User Data Service (UDS)	Directory Access for Unified Communications Endpoints Using Cisco User Data Service (UDS), page 16-6	March 1, 2018
Identity management	Identity Management Architecture Overview, page 16-33	March 1, 2018
Single Sign-On (SSO)	SSO for Cisco Jabber, page 16-43 Design Considerations for SSO, page 16-44	March 1, 2018
Authentication and OAuth 2.0	Authorization Framework, page 16-45	March 1, 2018

What is Directory Integration?

Integrating voice applications with a corporate LDAP directory is a common task for many enterprise IT organizations. However, the exact scope of the integration varies from company to company, and can translate into one or more specific and independent requirements, as shown in Figure 16-1.

Figure 16-1 Various Requirements for Directory Integration



One common requirement is to enable user lookups (sometimes called the "white pages" service) from IP phones or other voice and/or video endpoints, so that users can dial contacts quickly after looking up their numbers in the directory.

Another requirement is to provision users automatically from the corporate directory into the user database for applications. This method avoids having to add, remove, or modify core user information manually each time a change occurs in the corporate directory.

Authentication of end users and administrators of the voice and/or video applications using their corporate directory credentials is also a common requirement. Enabling directory authentication allows the IT department to deliver single log-on functionality while reducing the number of passwords each user needs to maintain across different corporate applications.

As shown in Table 16-2, within the context of a Cisco Unified Communications system, the term *directory access* refers to mechanisms and solutions that satisfy the requirement of user lookups for Cisco Unified Communications endpoints, while the term *directory integration* refers to mechanisms and solutions that satisfy the requirements of user provisioning and authentication (for both end users and administrators).

Table 16-2 Directory Requirements and Cisco Solutions

Requirement	Cisco Solution	Cisco Unified CM Feature
User lookup for endpoints	Directory access	Cisco Unified IP Phone Services SDK Cisco User Data Service (UDS)
User provisioning	Directory integration	LDAP Synchronization
Authentication for Unified Communications end users	Directory integration	LDAP Authentication
Authentication for Unified Communications application administrators	Directory integration	LDAP Authentication

The remainder of this chapter describes how to address these requirements in a Cisco Unified Communications system based on Cisco Unified CM.

Note Another interpretation of the term *directory integration* revolves around the ability to add application servers to a Microsoft Active Directory domain in order to centralize management and security policies. Cisco Unified CM is an appliance that runs on a customized embedded operating system, and it cannot be added to a Microsoft Active Directory domain. Server management for Unified CM is provided through the Cisco Real Time Monitoring Tool (RTMT). Strong security policies tailored to the application are already implemented within the embedded operating system.

Directory Access for Unified Communications Endpoints

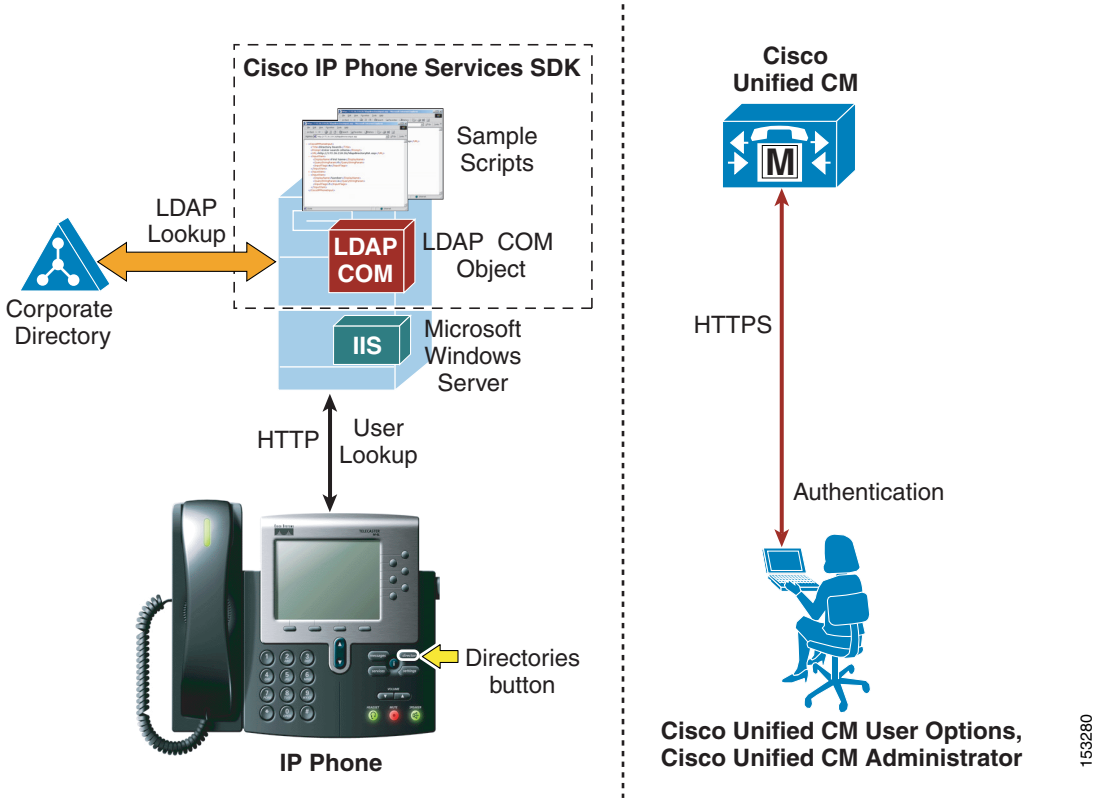
This section describes how to configure corporate directory access to any LDAP-compliant directory server to perform user lookups from Cisco Unified Communications endpoints (such as Cisco Unified IP Phones). The guidelines contained in this section apply regardless of whether Unified CM or other Unified Communications applications have been integrated with a corporate directory for user provisioning and authentication.

Cisco Unified IP Phones equipped with a display screen can search a user directory when a user presses the Directories button on the phone. The IP Phones use Hyper-Text Transfer Protocol (HTTP) to send requests to a web server. The responses from the web server contain specific Extensible Markup Language (XML) objects that the phone interprets and displays.

By default, Cisco Unified IP Phones are configured to perform user lookups against Unified CM's embedded database. However, it is possible to change this configuration so that the lookup is performed on a corporate LDAP directory. In this case, the phones send an HTTP request to an external web server that operates as a proxy by translating the request into an LDAP query which is then processed by the corporate directory. The web server encapsulates the LDAP response into an XML object that is sent back to the phone using HTTP, to be rendered to the end user.

Figure 16-2 illustrates this mechanism in a deployment where Unified CM has not been integrated with the corporate directory. Note that, in this scenario, Unified CM is not involved in the message exchange. The authentication mechanism to Unified CM web pages, shown on the right half of Figure 16-2, is independent of how directory lookup is configured.

Figure 16-2 Directory Access for Cisco Unified IP Phones Using the Cisco Unified IP Phone Services SDK



In the example shown in Figure 16-2, the web server proxy function is provided by the Cisco LDAP Search Component Object Model (COM) server, which is included in the Cisco Unified IP Phone Services Software Development Kit (SDK). You can download the latest Cisco Unified IP Phone Services SDK from Cisco DevNet, the Cisco developer community, at

https://developer.cisco.com/site/devnet/home/index.gsp

The IP Phone Services SDK can be installed on a Microsoft Windows web server running IIS 4.0 or later, but it cannot be installed on a Unified CM server. The SDK includes some sample scripts to provide simple directory lookup functionality.

To set up a corporate directory lookup service using the IP Phone Services SDK, perform the following steps:

Step 1 Modify one of the sample scripts to point to your corporate LDAP directory, or write your own script using the LDAP Search COM Programming Guide provided with the SDK.

Step 2 In Unified CM, configure the URL Directories parameter (under **System > Enterprise Parameters**) to point to the URL of the script on the external web server.

Step 3 Reset the phones to make the changes take effect.

Note If you want to offer the service only to a subset of users, configure the URL Directories parameter directly within the Phone Configuration page instead of the Enterprise Parameters page.

In conclusion, the following design considerations apply to directory access with the Cisco Unified IP Phone Services SDK:

- User lookups are supported against any LDAP-compliant corporate directory.
- When querying Microsoft Active Directory, you can perform lookups against the Global Catalog by pointing the script to a Global Catalog server and specifying port 3268 in the script configuration. This method typically results in faster lookups. Note that a Global Catalog does not contain a complete set of attributes for users. Refer to Microsoft Active Directory documentation for details.
- There is no impact on Unified CM when this functionality is enabled, and only minimal impact on the LDAP directory server.
- The sample scripts provided with the SDK allow only a minimal amount of customization (for example, you can prefix a digit string to all returned numbers). For a higher degree of manipulation, you will have to develop custom scripts, and a programming guide is included with the SDK to aid in writing the scripts.
- This functionality does not entail provisioning or authentication of Unified CM users with the corporate directory.

Directory Access for Unified Communications Endpoints Using Cisco User Data Service (UDS)

This section describes the mechanisms and best practices for directory access for endpoints using UDS to access user data instead of using the web service described in the previous section. The User Data Service (UDS) API is a REST-based set of operations that provide authenticated access to user resources and entities such as user devices, subscribed services, speed dials, and much more from the Unified Communications configuration database.

Current endpoints, including all endpoints running CE software, directly access the UDS REST-based directory search methods to obtain search results whenever a user invokes the search function on the endpoint. The results returned by the UDS search method are then displayed on the endpoint display. The UDS search function lists only those directory entries that exist in the Unified CM database, unless the UDS LDAP proxy functionality is used. When using the UDS LDAP proxy functionality, the endpoints still request directory information from Cisco Unified CM via UDS, but the results will then be requested by the UDS service from the configured external LDAP directory before being returned to the endpoint as a result of the UDS request.

Directory Integration with Unified CM

This section describes the mechanisms and best practices for directory integration with Cisco Unified CM to allow for user provisioning and authentication with a corporate LDAP directory. This section covers the following topics:

- Cisco Unified Communications Directory Architecture, page 16-7

  This section provides an overview of the user-related architecture in Unified CM.

- LDAP Synchronization, page 16-10

  This section describes the functionality of LDAP synchronization and provides design guidelines for its deployment, with additional considerations for Microsoft Active Directory.

- LDAP Authentication, page 16-22

  This section describes the functionality of LDAP authentication and provides design guidelines for its deployment, with additional considerations for Microsoft Active Directory.

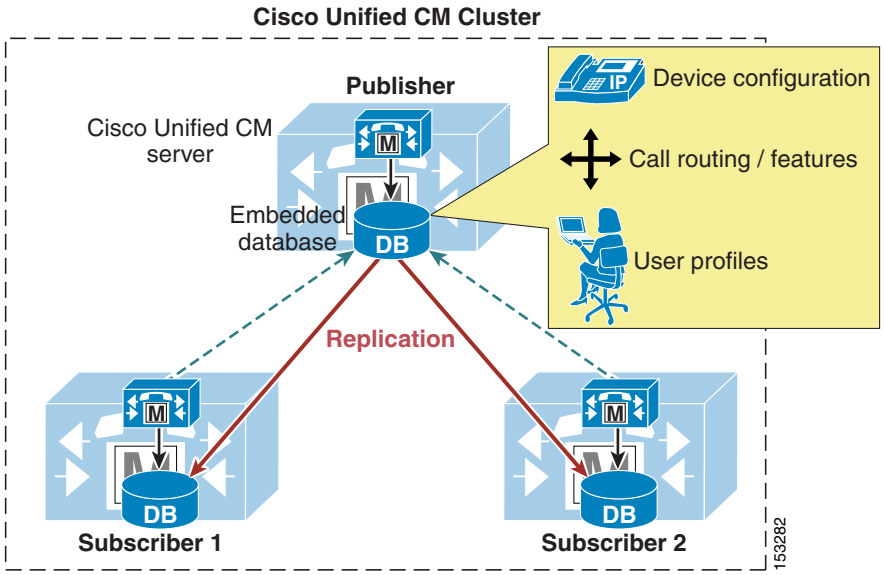
For a list of supported LDAP directories, refer to the latest version of the *System Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Cisco Unified Communications Directory Architecture

Figure 16-3 shows the basic architecture of a Unified CM cluster. The embedded database stores all configuration information, including device-related data, call routing, feature provisioning, and user profiles. The database is present on all servers within a Unified CM cluster and is replicated automatically from the publisher server to all subscriber servers.

Figure 16-3 Cisco Unified CM Architecture



By default, all users are provisioned manually in the publisher database through the Unified CM Administration web interface. Cisco Unified CM has two types of users:

- End users — All users associated with a physical person and an interactive login. This category includes all Unified Communications users as well as Unified CM administrators when using the User Groups and Roles configuration (equivalent to the Cisco Multilevel Administration feature in prior Unified CM versions).
- Application users — All users associated with other Cisco Unified Communications features or applications, such as Cisco Attendant Console, Cisco Unified Contact Center Express, or Cisco Unified Communications Manager Assistant. These applications need to authenticate with Unified CM, but these internal "users" do not have an interactive login and serve purely for internal communications between applications.

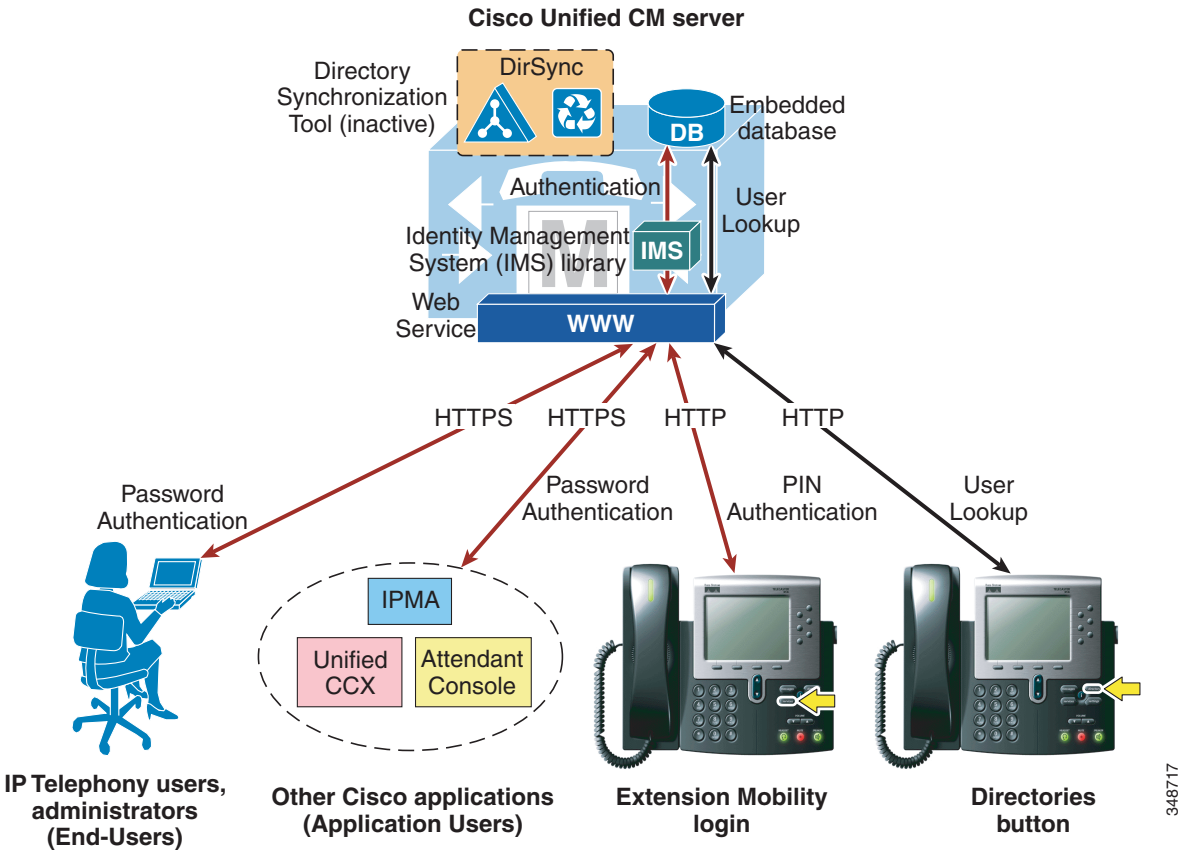
Table 16-3 lists the application users created by default in the Unified CM database, together with the feature or application that uses them. Additional application users can be created manually when integrating other Cisco Unified Communications applications (for example, the **ac** application user for Cisco Attendant Console, the **jtapi** application user for Cisco Unified Contact Center Express, and so forth).

Table 16-3 Default Application Users for Unified CM

Application User	Used by:
CCMAdministrator	Unified CM Administration (default "super user")
CCMQRTSecureSysUser	Cisco Quality Reporting Tool
CCMQRTSysUser	
CCMSysUser	Cisco Extension Mobility
IPMASecureSysUser	Cisco Unified Communications Manager Assistant
IPMASysUser	
WDSecureSysUser	Cisco WebDialer
WDSysUser	

Based on these considerations, Figure 16-4 illustrates the default behavior in Unified CM for user-related operations such as lookups, provisioning, and authentication.

Figure 16-4 Default Behavior for User-Related Operations for Unified CM



End users access the Unified CM User Options page via HTTPS and authenticate with a user name and password. If they have been configured as administrators by means of User Groups and Roles, they can also access the Unified CM Administration pages with the same credentials.

Similarly, other Cisco features and applications authenticate to Unified CM via HTTPS with the user name and password associated with their respective application users.

The authentication challenge carried by the HTTPS messages are relayed by the web service on Unified CM to an internal library called Identity Management System (IMS). In its default configuration, the IMS library authenticates both end users and application users against the embedded database. In this way, both "physical" users of the Unified Communications system and internal application accounts are authenticated using the credentials configured in Unified CM.

End users may also authenticate with their user name and a numeric password (or PIN) when logging into the Extension Mobility service from an IP phone. In this case, the authentication challenge is carried via HTTP to Unified CM but is still relayed by the web service to the IMS library, which authenticates the credentials against the embedded database.

In addition, user lookups performed by Unified Communications endpoints via the Directories button communicate with the web service on Unified CM via HTTP and access data on the embedded database.

The importance of the distinction between End Users and Application Users becomes apparent when integration with a corporate directory is required. As mentioned in the previous section, this integration is accomplished by means of the following two separate processes:

- LDAP synchronization

  This process uses an internal tool called Cisco Directory Synchronization (DirSync) on Unified CM to synchronize a number of user attributes (either manually or periodically) from a corporate LDAP directory. When this feature is enabled, users are automatically provisioned from the corporate directory in addition to local user provisioning through the Unified CM administration GUI. This feature applies only to End Users, while Application Users are kept separate and are still provisioned via the Unified CM Administration interface. In summary, End Users are defined in the corporate directory and synchronized into the Unified CM database, while Application Users are stored only in the Unified CM database and do not need to be defined in the corporate directory.

- LDAP authentication

  This process enables the IMS library to authenticate user credentials of LDAP synchronized End Users against a corporate LDAP directory using the LDAP standard Simple_Bind operation. When this feature is enabled, End User passwords of LDAP synchronized End Users are authenticated against the corporate directory, while Application User passwords and passwords of local End Users are still authenticated locally against the Unified CM database. Cisco Extension Mobility PINs are also still authenticated locally.

Maintaining and authenticating the Application Users internally to the Unified CM database provides resilience for all the applications and features that use these accounts to communicate with Unified CM, independently of the availability of the corporate LDAP directory.

Cisco Extension Mobility PINs are also kept within the Unified CM database because they are an integral part of a real-time application, which should not have dependencies on the responsiveness of the corporate directory.

The next two sections describe in more detail LDAP synchronization and LDAP authentication, and they provide design best-practices for both functions.

Note As illustrated in the section on Directory Access for Unified Communications Endpoints, page 16-4, user lookups from endpoints can also be performed against a corporate directory by configuring the Cisco Unified IP Phone Services SDK on an external web server.

LDAP Synchronization

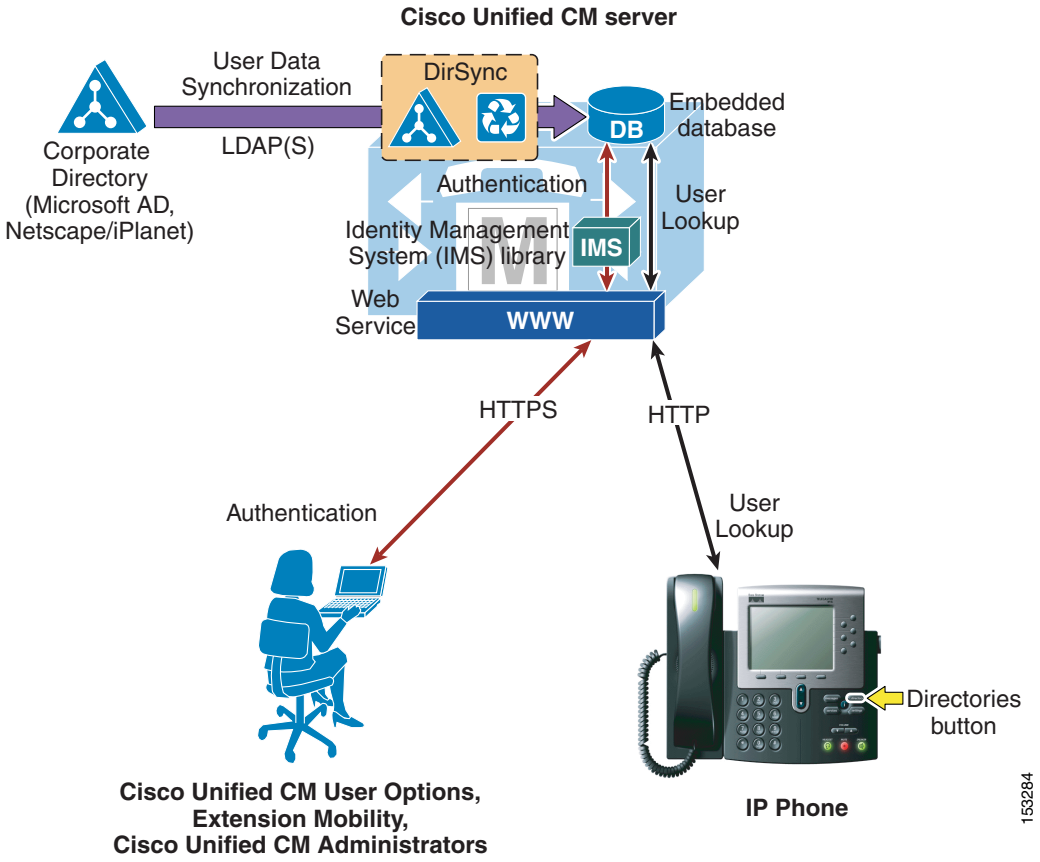
Synchronization of Unified CM with a corporate LDAP directory allows the administrator to provision users easily by mapping Unified CM data fields to directory attributes. Critical user data maintained in the LDAP store is copied into the appropriate corresponding fields in the Unified CM database on a scheduled or on-demand basis. The corporate LDAP directory retains its status as the central repository. Unified CM has an integrated database for storing user data and a web interface within Unified CM Administration for creating and managing user accounts and data. When LDAP synchronization is enabled, the local database is still used, and additional local end-user accounts can be created. Management of end-user accounts is then accomplished through the interface of the LDAP directory and the Unified CM administration GUI. (See Figure 16-5.). Accounts for application users can be created and managed only through the Unified CM Administration web interface.

The user account information is imported from the LDAP directory into the database located on the Unified CM publisher server. Information that is imported from the LDAP directory may not be changed by Unified CM. Additional user information specific to Cisco Unified Communications is managed by

Unified CM and stored only within its local database. For example, device-to-user associations, speed dials, call forward settings, and user PINs are all examples of data that is managed by Unified CM and does not exist in the corporate LDAP directory. The user data is then propagated from the Unified CM publisher server to the subscriber servers through the built-in database synchronization mechanism.

User information synchronized from the LDAP directory can be converted to local user information so that the user information then can be edited locally on Unified CM. Local end users can be added manually using the Unified CM administration GUI. During an LDAP sync, a local end user is converted to an active LDAP user, and if a user with the same user ID is found in LDAP, the locally configured data is replaced with data from the directory.

Figure 16-5 Enabling Synchronization of User Data



When LDAP synchronization is activated, only one type of LDAP directory may be chosen globally for the cluster at any one time. Also, one attribute of the LDAP directory user is chosen to map into the Unified CM User ID field. Unified CM uses standard LDAPv3 for accessing the data.

Cisco Unified CM imports data from standard attributes. Extending the directory schema is not required. Table 16-4 lists the attributes that are available for mapping to Unified CM fields. The data of the directory attribute that is mapped to the Unified CM User ID must be unique within all entries for that cluster. The attribute mapped to the Unified CM UserID field must be populated in the directory and the **sn** attribute must be populated with data, otherwise those records are skipped during this import action. If the primary attribute used during import of end-user accounts matches any application user in the Unified CM database, that user is not imported from the LDAP directory.

Table 16-4 lists the attributes that are imported from the LDAP directory into corresponding Unified CM user fields, and it describes the mapping between those fields. Some Unified CM user fields might be mapped from one of several LDAP attributes.

Table 16-4 Synchronized LDAP Attributes and Corresponding Unified CM Field Names

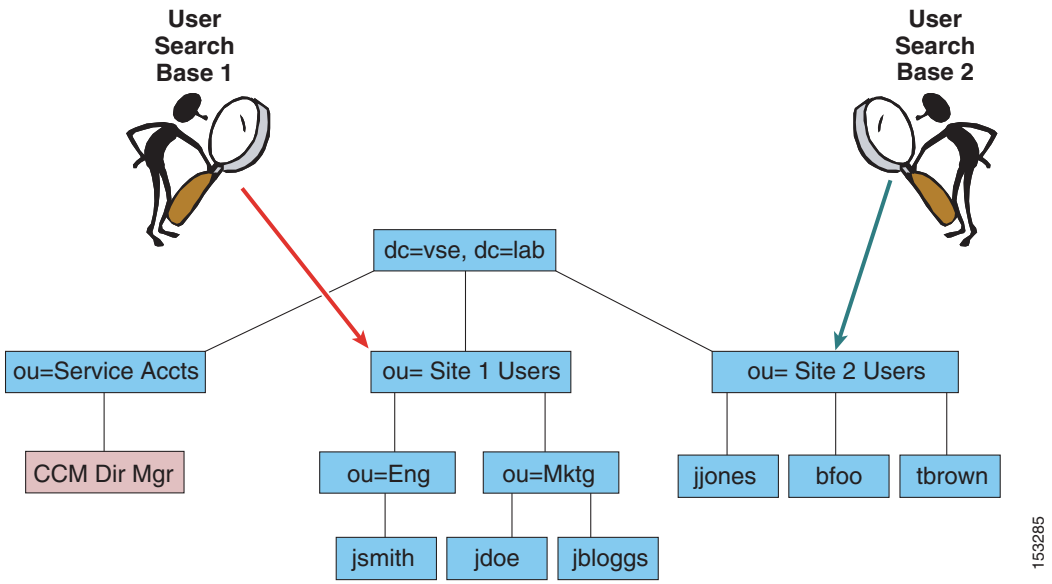
Unified CM User Field	Microsoft Active Directory	Microsoft Active Directory Application Mode (ADAM) or Active Directory Lightweight Directory Service (AD LDS)	Oracle DSEE and Sun	OpenLDAP and Other LDAPv3 Types
User ID	*One of:* sAMAccountName mail employeeNumber telephoneNumber userPrincipalName	*One of:* uid mail employeeNumber telephoneNumber userPrincipalName	*One of:* uid mail employeeNumber telephonePhone	*One of:* uid mail employeeNumber telephonePhone
First Name	givenName	givenName	givenName	givenName
Middle Name	*One of:* middleName initials	*One of:* middleName initials	initials	initials
Last Name	sn	sn	sn	sn
Manager ID	manager	manager	manager	manager
Department	department	department	departmentnumber	departmentnumber
Phone Number	*One of:* telephoneNumber ipPhone	*One of:* telephoneNumber ipPhone	telephoner	telephonenumber
Mail ID	*One of:* mail sAMAccountName	*One of:* mail uid	*One of:* mail uid	*One of:* mail uid
objectGUID	objectGUID	objectGUID	not applicable	not applicable
OCSPrimaryUser Address	msRTCSIP-PrimaryUser Address	not applicable	not applicable	not applicable
Title	title	title	Title	title
Home Phone Number	homePhone	homePhone	Homephone	hometelephonenumber
Mobile Phone Number	mobile	mobile	Mobile	Mobiletelephonenumber
Pager Number	pager	pager	Pager	Pagertelephonenumber
Directory URI	*One of:* msRTCSIP-PrimaryUser Address mail none	*One of:* mail none	*One of:* mail none	*One of:* mail none
Display Name	displayName	displayName	displayName	displayName

In addition to the direct mapping of directory attributes to local user attributes, other characteristics of the synchronized users are determined by settings on the LDAP directory synchronization agreement. Access control group membership of users created through LDAP synchronization is directly configured in the LDAP directory configuration setting. Further user capabilities are determined by the feature group template selected. The selection of a feature group template on an LDAP directory synchronization agreement is optional. The feature group templates allow administrators to define user characteristics, including home cluster selection, IM and Presence capabilities, mobility features, services profiles, and user profiles. The user profiles allow administrators to define a universal line template that is considered for automatic creation of directory numbers for LDAP synchronized users by Unified CM.

The synchronization is performed by a process called Cisco DirSync, which is enabled through the Serviceability web page. When enabled, it allows one to 20 synchronization agreements to be configured in the system. This number is reduced to 10 if more than 80,000 users are synchronized. An agreement specifies a search base that is a position in the LDAP tree where Unified CM will begin its search for user accounts to import. Unified CM can import only users that exist in the domain specified by the search base for a particular synchronization agreement.

In Figure 16-6, two synchronization agreements are represented. One synchronization agreement specifies User Search Base 1 and imports users jsmith, jdoe and jbloggs. The other synchronization agreement specifies User Search Base 2 and imports users jjones, bfoo, and tbrown. The CCMDirMgr account is not imported because it does not reside below the point specified by a user search base. When users are organized in a structure in the LDAP directory, you can use that structure to control which user groups are imported. In this example, a single synchronization agreement could have been used to specify the root of the domain, but that search base would also have imported the Service Accts. The search base does not have to specify the domain root; it may specify any point in the tree.

Figure 16-6 User Search Bases



To import the data into the Unified CM database, the system performs a bind to the LDAP directory using the account specified in the configuration as the LDAP Manager Distinguished Name, and reading of the database is done with this account. The account must be available in the LDAP directory for Unified CM to log in, and Cisco recommends that you create a specific account with permissions to

allow it to read all user objects within the sub-tree that was specified by the user search base. The sync agreement specifies the full Distinguished Name of that account so that the account may reside anywhere within that domain. In the example in Figure 16-6, CCMDirMgr is the account used for the synchronization.

It is possible to control the import of accounts through use of permissions of the LDAP Manager Distinguished Name account. In this example, if that account is restricted to have read access to ou=Eng but not to ou=Mktg, then only the accounts located under Eng will be imported.

Synchronization agreements have the ability to specify multiple directory servers to provide redundancy. You can specify an ordered list of up to three directory servers in the configuration that will be used when attempting to synchronize. The servers are tried in order until the list is exhausted. If none of the directory servers responds, then the synchronization fails, but it will be attempted again according to the configured synchronization schedule.

Synchronization Mechanism

The synchronization agreement specifies a time for synchronizing to begin and a period for re-synchronizing that can be specified in hours, days, weeks, or months (with a minimum value of 6 hours). A synchronization agreement can also be set up to run only once at a specific time.

When synchronization is enabled for the first time on a Unified CM publisher server, user accounts that exist in the corporate directory are imported into the Unified CM database. Then either existing Unified CM end-user accounts are activated and data is updated, or a new end-user account is created according to the following process:

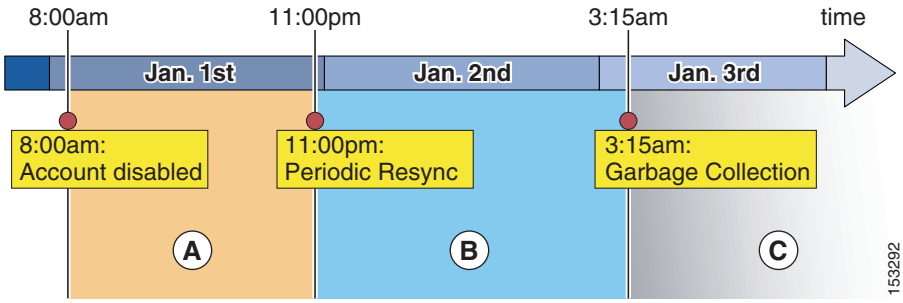
1. If end-user accounts already exist in the Unified CM database and a synchronization agreement is configured, all pre-existing accounts that have been synchronized from LDAP previously are marked inactive in Unified CM. The configuration of the synchronization agreement specifies a mapping of an LDAP database attribute to the Unified CM UserID. During the synchronization, accounts from the LDAP database that match an existing Unified CM account cause that Unified CM account to be marked active again.
2. After the synchronization is completed, any LDAP synchronized accounts that were not set to active are permanently deleted from Unified CM when the garbage collection process runs. Garbage collection is a process that runs automatically at the fixed time of 3:15 AM, and it is not configurable.
3. Subsequently when changes are made in the corporate directory, the synchronization from Microsoft Active Directory occurs as a full re-synchronization at the next scheduled synchronization period. On the other hand, the Sun ONE directory products perform an incremental synchronization triggered by a change in the directory. The following sections present examples of each of these two scenarios.

Note Once users are synchronized from LDAP into the Unified CM database, deletion of a synchronization configuration will cause users that were imported by that configuration to be marked inactive in the database. Garbage collection will subsequently remove those users.

Account Synchronization with Active Directory

Figure 16-7 shows an example timeline of events for a Unified CM deployment where LDAP Synchronization and LDAP Authentication have both been enabled. The re-synchronization is set for 11:00 PM daily.

Figure 16-7 Change Propagation with Active Directory



After the initial synchronization, the creation, deletion, or disablement of an account will propagate to Unified CM according to the timeline shown in Figure 16-7 and as described in the following steps:

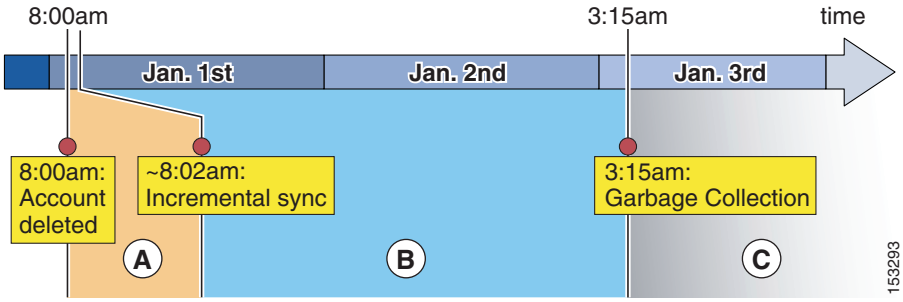
1. At 8:00 AM on January 1, an account is disabled or deleted in AD. From this time and during the whole period A, password authentication (for example, Unified CM User Options page) will fail for this user because Unified CM redirects authentication to AD. However, PIN authentication (for example, Extension Mobility login) will still succeed because the PIN is stored in the Unified CM database.
2. The periodic re-synchronization is scheduled for 11:00 PM on January 1. During that process, Unified CM will verify all accounts. Any accounts that have been disabled or deleted from AD will at that time be tagged in the Unified CM database as inactive. After 11:00 PM on January 1, when the account is marked inactive, both the PIN and password authentication by Unified CM will fail.
3. Garbage collection of accounts occurs daily at the fixed time of 3:15 AM. This process permanently deletes user information from the Unified CM database for any record that has been marked inactive for over 24 hours. In this example, the garbage collection that runs at 3:15 AM on January 2 does not delete the account because it has not been inactive for 24 hours yet, so the account is deleted at 3:15 AM on January 3. At that point, the user data is permanently deleted from Unified CM.

If an account has been created in AD at the beginning of period A, it will be imported to Unified CM at the periodic re-synchronization that occurs at the beginning of period B and will immediately be active on Unified CM.

Account Synchronization with Sun ONE

Sun ONE products support incremental synchronization agreements and use a different synchronization timeline than Microsoft Active Directory. The synchronization makes use of the Persistent Search mechanism supported by many LDAP implementations. Figure 16-8 shows an example of this synchronization timeline for a Unified CM deployment with LDAP Synchronization and LDAP Authentication both enabled.

Figure 16-8 Change Propagation with Sun ONE



The example in Figure 16-8 involves the following steps:

1. An account is deleted from the corporate directory at 8:00 AM on January 1, which causes an incremental update to be sent from the LDAP server to Unified CM. Unified CM sets its corresponding copy of the data to inactive. Because LDAP authentication is configured, the user will be unable to log in via password as soon as the LDAP server has deleted the record. Also, the PIN may not be used for login at the moment the Unified CM record is marked inactive.
2. During period B, the user's record is still present in Unified CM, albeit inactive.
3. When the garbage collection runs at 3:15 AM on January 2, the record has not yet been inactive for 24 hours. The data remains in the Unified CM database until the beginning of period C on January 3, when the garbage collection process runs again at 3:15 AM and determines that the record has been inactive for 24 hours or more. The record is then permanently deleted from the database.

Accounts that are newly created in the directory are synchronized to Unified CM via incremental updates as well, and they may be used as soon as the incremental update is received.

Automatic Line Creation

For users created during LDAP synchronization, Unified CM can automatically create directory numbers. These auto-generated directory numbers are either based on information found in the directory and defined based on a mask to be applied to the phone number found in the directory, or the numbers are taken from directory number pools defined on the LDAP synchronization agreement. If a mask is defined on the synchronization agreement, then to allow for variable length +E.164 directory numbers to be generated, the following rules apply:

- If the mask is left empty, then Unified CM takes all digits and also a leading "+" (if present) from the directory.
- X is used as a wildcard character in the mask.
- A wildcard matches on digits and "+".
- Wildcards in the mask are filled from the right.
- Unfilled wildcards in the mask are removed.

Table 16-5 shows some examples.

Table 16-5 Examples for Directory Number Creation from LDAP Phone Numbers Based on Masks

Number in LDAP	Mask	Result
14085551234		14085551234
14085551234	+XXXXXXXXXX	+14085551234
14085551234	+XXXXXXXXXXXXXXXXX	+14085551234
14085551234	XXXX	1234
+14085551234		+14085551234
+14085551234	+XXXXXXXXXXXXXXXXX	+14085551234
+496100123	+XXXXXXXXXXXXXXXXX	+496100123

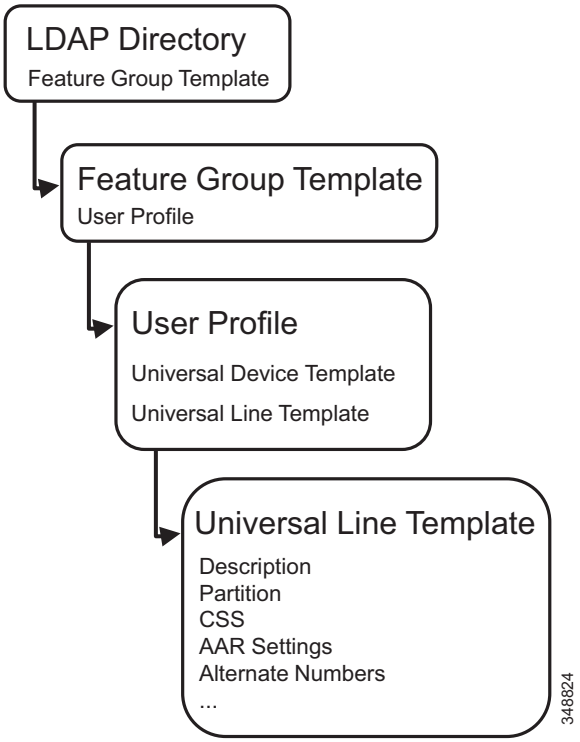
As an alternative to creating directory numbers based on information from LDAP, directory numbers for new users can also be taken from predefined number pools. Each pool is defined by a start and end number. Directory number pools support +E.164 numbers. Up to five pools can be defined. Numbers are assigned from the first pool until all numbers of that pools have been assigned. Number assignment then starts to take numbers from the next pool.

Automatic Line Creation is enabled only if *both* of the following conditions are met:

- A Feature Group Template is assigned in the directory synchronization agreement, **and**
- A Universal Line Template is selected in the User Profile selected in the Feature Group Template.

Figure 16-9 shows the hierarchy of configuration elements required to define line-level settings for automatic line creation.

Figure 16-9 Relation of LDAP Directory Configuration, Feature Group Template, User Profile, and Universal Line Template



Ultimately the Universal Line Template defines the characteristics for all directory numbers that are automatically created for users added through the corresponding LDAP synchronization definition.

Design Considerations

The calling search space defined in the Universal Line Template determines the class of service of devices using any of the auto-generated directory numbers. This implies that all directory numbers created through the same LDAP synchronization agreement share the same class of service, and thus if directory numbers for multiple sites and multiple classes of service need to be auto-generated, then multiple LDAP synchronization agreements (one per site and class of service) need to be configured. For each of these synchronization agreements, disjunct LDAP filters need to be defined, each exactly matching on only the users belonging to one of the site-specific and class-of-service-specific user groups. This mapping from LDAP attributes to site and class of service groups can be challenging unless the group membership based on site and class of service is explicitly encoded in few LDAP attributes (potentially even in a custom attribute). Also, the maximum number of supported LDAP agreements is limited, which limits the number of distinct user groups for which directory numbers can be created automatically.

Automatic creation of directory numbers applies only to users created during LDAP directory synchronization. Adding, changing, or updating the Universal Line Template for a given LDAP synchronization agreement will not create directory numbers for already existing users and will not change the settings of already existing directory numbers.

The Universal Line Template allows administrators to define call forward unregistered destinations and either to select voicemail as the forward destination or to define an explicit destination. To reach endpoints in remote sites from registered endpoints in case of WAN failure, the call forward unregistered destination for the remote site's phones must be set to the PSTN alias (+E.164 number) of the remote

phone. This cannot be achieved with Universal Line Template settings because this would require defining the call forward unregistered destination to be set based on the assigned directory numbers (potentially with a mask applied).

Enterprise Group Support

To enable Jabber clients to search for groups in Microsoft Active Directory, you can configure Unified CM not only to synchronize end users from Active Directory but also to include distribution groups defined in Active Directory. Synchronization of enterprise groups is supported only with Microsoft Active Directory as the data source. It is not supported with Active Directory Lightweight Directory Services (AD LDS) or other corporate directories. Synchronization of enterprise groups is enabled in the Unified CM LDAP directory configuration. The maximum number of enterprise groups is 15,000 and the maximum number of members per group is 100. While groups and members cannot be added or modified in the Unified CM administration, the groups synchronized from Active Directory can be reviewed in the User Management/User Settings/User Group menu.

For each group member, the following information is available on Jabber clients:

- Display name
- User ID
- Title
- Phone number
- Mail ID

Security Considerations

During the import of accounts, no passwords or PINs are copied from the LDAP directory to the Unified CM database. If LDAP authentication is not enabled in Unified CM and single sign-on is not used, the password for the end user is managed by using Unified CM Administration. The password and PIN are stored in an encrypted format in the Unified CM database. The PIN is always managed on Unified CM. If you want to use the LDAP directory password to authenticate an end user, see the section on LDAP Authentication, page 16-22.

The connection between the Unified CM publisher server and the directory server can be secured by enabling Secure LDAP (SLDAP) on Unified CM and the LDAP server. Secure LDAP enables LDAP to be sent over a Secure Socket Layer (SSL) connection and can be enabled by adding the LDAP server into the Tomcat trust store within the Unified CM Platform Administration. For detailed procedure steps, refer to the Unified CM product documentation available at https://www.cisco.com. Refer to the documentation of the LDAP directory vendor to determine how to enable SLDAP.

Design Considerations for LDAP Synchronization

Observe the following design and implementation best practices when deploying LDAP synchronization with Cisco Unified CM:

- Use a specific account within the corporate directory to allow the Unified CM synchronization agreement to connect and authenticate to it. Cisco recommends that you use an account dedicated to Unified CM, with minimum permissions set to "read" all user objects within the desired search base and with a password set never to expire. The password for this account in the directory must be kept in synchronization with the password configuration of the account in Unified CM. If the service account password changes in the directory, be sure to update the account configuration in Unified CM.

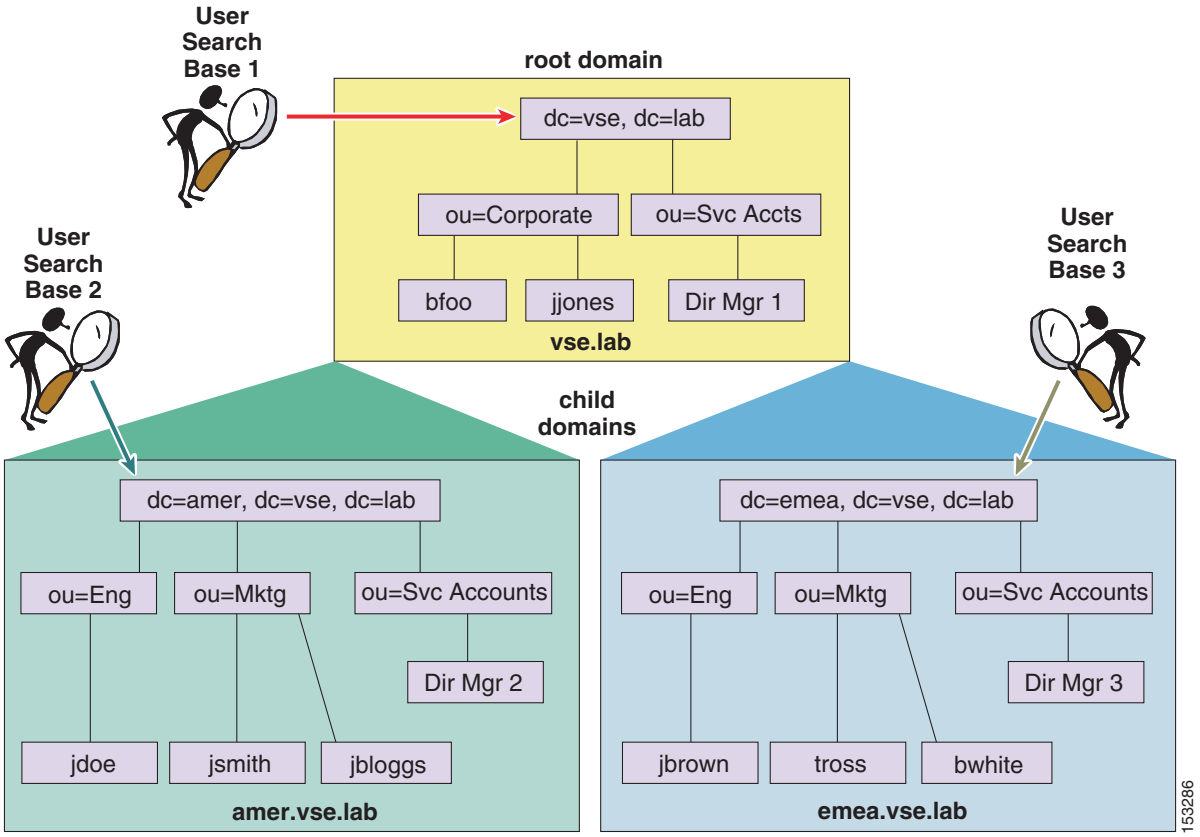
- All synchronization agreements on a given cluster must integrate with the same family of LDAP servers.
- Stagger the scheduling of synchronization agreements so that multiple agreements are not querying the same LDAP servers simultaneously. Choose synchronization times that occur during quiet periods (off-peak hours).
- If security of user data is required, enable Secure LDAP (SLDAP) by checking the **Use SSL** field on the LDAP Directory configuration page in Unified CM Administration.
- Ensure that the LDAP directory attribute chosen to map into the Unified CM UserID field is unique within all synchronization agreements for that cluster.
- The attribute chosen as UserID must not be the same as that for any of the Application Users defined in Unified CM.
- The LDAP attribute sn(lastname) is a mandatory attribute for LDAP Synchronization of users.
- An existing account in the Unified CM database before synchronization is maintained only if an account imported from the LDAP directory has a matching attribute. The attribute that is matched to the Unified CM UserID is determined by the synchronization agreement.
- Administer end-user accounts through the LDAP directory's management tools, and manage the Cisco-specific data for those accounts through the Unified CM Administration web page.
- For AD deployments, the ObjectGUID is used internally in Unified CM as the key attribute of a user. The attribute in AD that corresponds to the Unified CM User ID may be changed in AD. For example, if sAMAccountname is being used, a user may change their sAMAccountname in AD, and the corresponding user record in Unified CM would be updated.

  With all other LDAP platforms, the attribute that is mapped to User ID is the key for that account in Unified CM. Changing that attribute in LDAP will result in a new user being created in Unified CM, and the original user will be marked inactive.

Additional Considerations for Microsoft Active Directory

A synchronization agreement for a domain will not synchronize users outside of that domain nor within a child domain because Unified CM does not follow AD referrals during the synchronization process. The example in Figure 16-10 requires three synchronization agreements to import all of the users. Although Search Base 1 specifies the root of the tree, it will not import users that exist in either of the child domains. Its scope is only VSE.LAB, and separate agreements are configured for the other two domains to import those users.

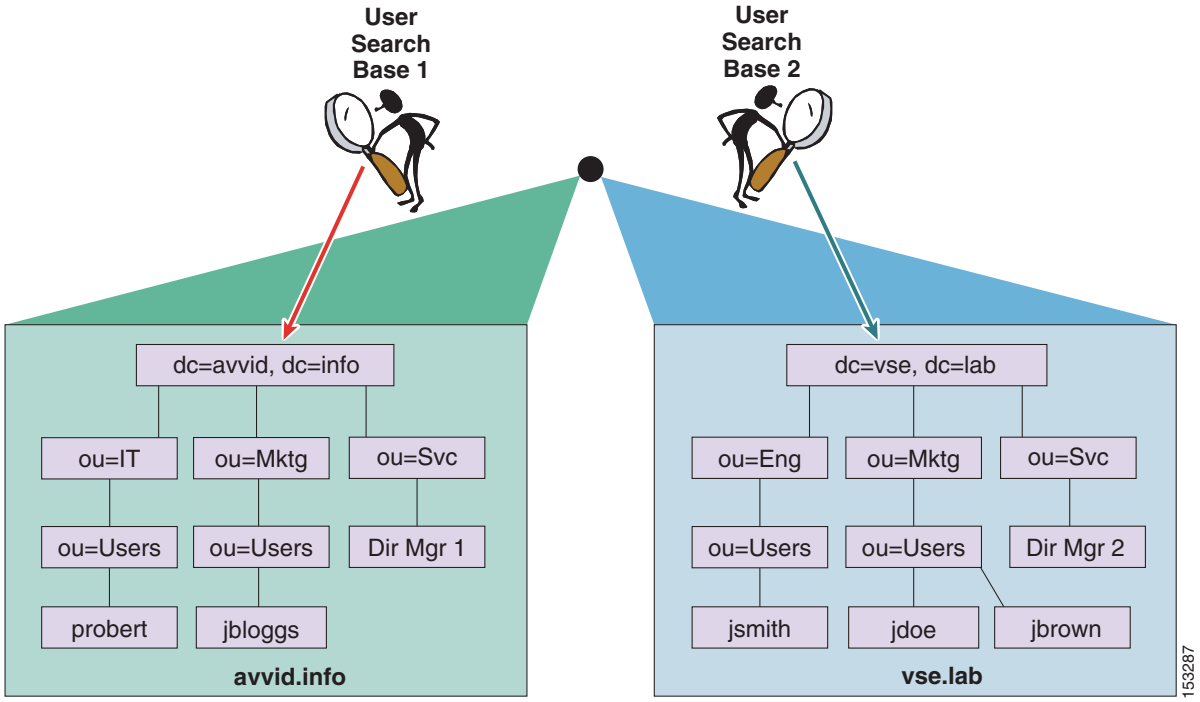
Figure 16-10 Synchronization with Multiple Active Directory Domains



In Figure 16-10, each of the domains and sub-domains contains at least one domain controller (DC) associated to them, and the three synchronization agreements each specify the appropriate domain controller. The DCs have information only on users within the domain where they reside, therefore three synchronization agreements are required to import all of the users.

When synchronization is enabled with an AD forest containing multiple trees, as shown in Figure 16-11, multiple synchronization agreements are still needed for the same reasons listed above. Additionally, the UserPrincipalName (UPN) attribute is guaranteed by Active Directory to be unique across the forest and must be chosen as the attribute that is mapped to the Unified CM UserID. For additional considerations on the use of the UPN attribute in a multi-tree AD scenario, see the section on Additional Considerations for Microsoft Active Directory, page 16-26.

Figure 16-11 Synchronization with Multiple AD Trees (Discontiguous Namespaces)



Unified CM sends a default LDAP search filter string to AD when performing the synchronization of accounts. One of the clauses is to not return accounts that have been marked as disabled in AD. An account marked disabled by AD, such as when failed login attempts are exceeded, will be marked inactive if synchronization runs while the account is disabled.

Unified CM Multi-Forest LDAP Synchronization

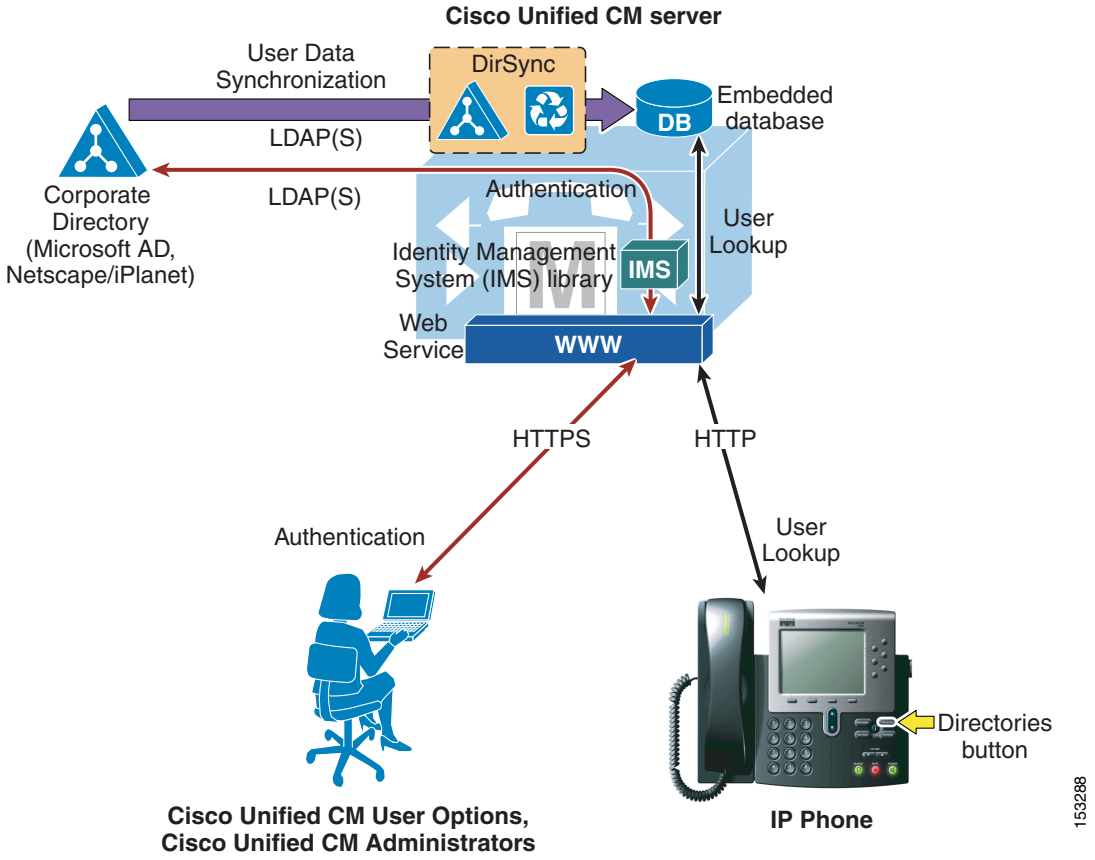
A Unified CM deployment using a multi-forest LDAP infrastructure can be supported by using Active Directory Lightweight Directory Services (AD LDS) as a single forest view integrating with the multiple disparate forests. The integration also requires the use of LDAP filtering (see User Filtering for Directory Synchronization and Authentication, page 16-28). For full details, refer to the document on *How to Configure Unified Communication Manager Directory Integration in a Multi-Forest Environment*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps556/products_configuration_example09186a0080b2b103.shtml

LDAP Authentication

The LDAP authentication feature enables Unified CM to authenticate LDAP synchronized users against the corporate LDAP directory. Application users and locally configured users are always authenticated against the local database. Also PINs of all end users are always checked against the local database only. This authentication is accomplished with an LDAPv3 connection established between the Identity Management System (IMS) module within Unified CM and a corporate directory server, as shown in Figure 16-12.

Figure 16-12 Enabling LDAP Authentication



To enable authentication, a single authentication agreement may be defined for the entire cluster. The authentication agreement supports configuration of up to three LDAP servers for redundancy and also supports secure connections LDAP over SSL (SLDAP) if desired. Authentication can be enabled only when LDAP synchronization is properly configured and used. LDAP authentication configuration is overridden by enabling SSO. With SSO enabled, end users are always authenticated using SSO, and LDAP authentication configuration is ignored.

The following statements describe Unified CM's behavior when authentication is enabled:

- End user passwords of users imported from LDAP are authenticated against the corporate directory by a simple bind operation.
- End user passwords for local users are authenticated against the Unified CM database.
- Application user passwords are authenticated against the Unified CM database.
- End user PINs are authenticated against the Unified CM database.

This behavior is in line with the guiding principle of providing single logon functionality for end users while making the operation of the real-time Unified Communications system independent of the availability of the corporate directory, and is shown graphically in Figure 16-13.

Figure 16-13 Authenticating End User Passwords, Application User Passwords, and End User PINs

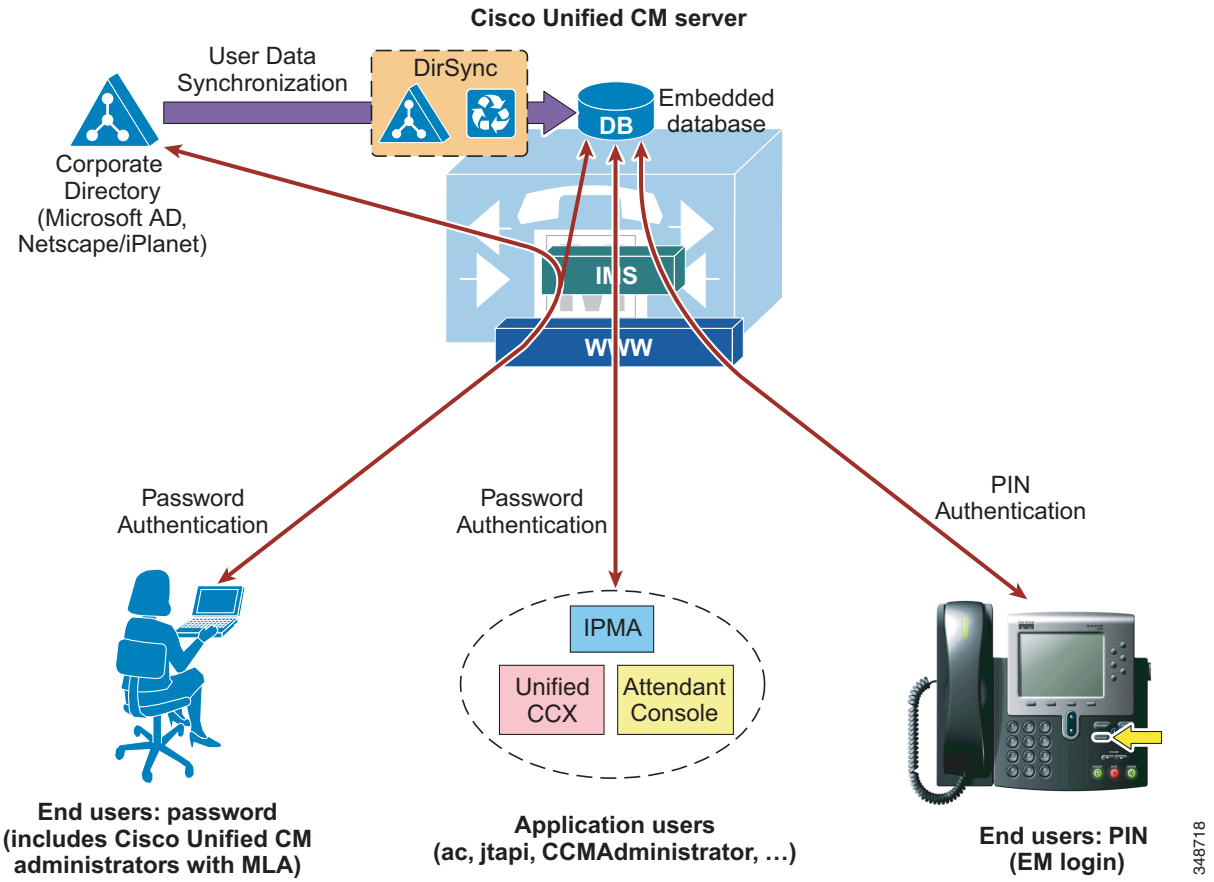


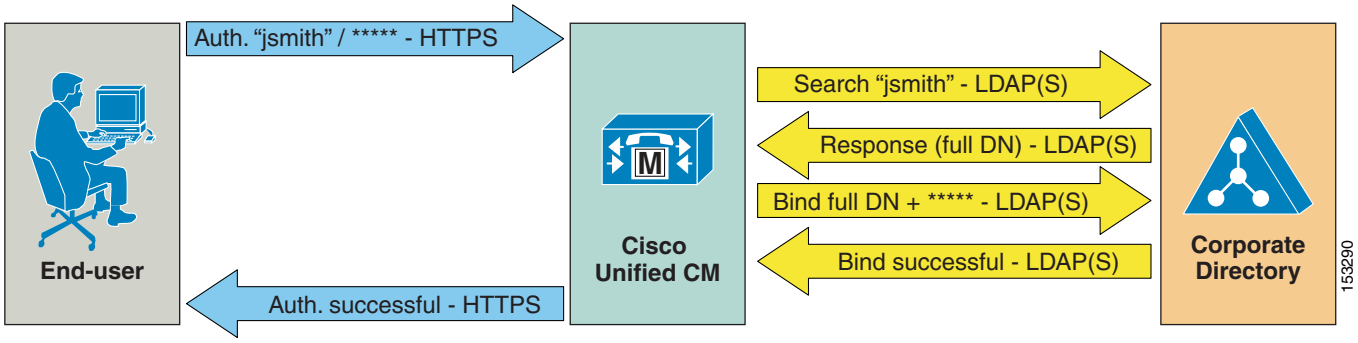
Figure 16-14 illustrates the following process, adopted by Unified CM to authenticate an end user synchronized from LDAP against a corporate LDAP directory:

1. A user connects to the Unified CM User Options page via HTTPS and attempts to authenticate with a user name and password. In this example, the user name is jsmith.
2. If the user is a local user, the password is checked against the local database.

The following steps apply only to LDAP synchronized users:

3. If the user is an LDAP synchronized user, Unified CM issues an LDAP query for the user name jsmith, using the value specified in the LDAP Search Base on the LDAP Authentication configuration page as the scope for this query. If SLDAP is enabled, this query travels over an SSL connection.
4. The corporate directory server replies via LDAP with the full Distinguished Name (DN) of user jsmith (for example, "cn=jsmith, ou=Users, dc=vse, dc=lab").
5. Unified CM then attempts to validate the user's credentials by using an LDAP bind operation to pass the full DN and password provided by the user.
6. If the LDAP bind is successful, Unified CM allows the user to proceed to the configuration page requested.

Figure 16-14 Authentication Process



Design Considerations for LDAP Authentication

Observe the following design and implementation best-practices when deploying LDAP authentication with Cisco Unified CM:

- Create a specific account within the corporate directory to allow Unified CM to connect and authenticate to it. Cisco recommends that you use an account dedicated to Unified CM, with minimum permissions set to "read" all user objects within the desired search base and with a password set to never expire. The password for this account in the directory must be kept in synchronization with the password configuration of the account in Unified CM. If the account password changes in the directory, be sure to update the account configuration in Unified CM. If LDAP synchronization is also enabled, you can use the same account for both functions.
- Enable LDAP authentication on Unified CM by specifying the credentials of the aforementioned account under LDAP Manager Distinguished Name and LDAP Password, and by specifying the directory subtree where all the users reside under LDAP User Search Base.
- This method provides single logon functionality to all end users synchronized from LDAP. They can then use their corporate directory credentials to log in to the Unified CM User Options page.
- Manage end-user passwords for LDAP synchronized users from within the corporate directory interface. Note that the password field is no longer displayed for LDAP synchronized users in the Unified CM Administration pages when authentication is enabled.
- Manage end-user PINs from the Unified CM Administration web pages or from the Unified CM User Options page.
- Manage Application User passwords from the Unified CM Administration web pages. Remember that these application users facilitate communication and remote call control with other Cisco Unified Communications applications and are not associated with real people.
- Enable single logon for Unified CM administrators by adding their corresponding end user to the Unified CM Super Users user group from the Unified CM Administration web pages. Multiple levels of administrator rights can be defined by creating customized user groups and roles.

Additional Considerations for Microsoft Active Directory

In environments that employ a distributed AD topology with multiple domain controllers geographically distributed, authentication speed might be unacceptable. When the Domain Controller for the authentication agreement does not contain a user account, a search must occur for that user across other domain controllers. If this configuration applies, and login speed is unacceptable, it is possible to set the authentication configuration to use a Global Catalog Server.

An important restriction exists, however. A Global Catalog does not carry the employeeNumber attribute by default. In that case either use Domain Controllers for authentication (beware of the limitations listed above) or update the Global Catalog to include the employeeNumber attribute. Refer to Microsoft Active Directory documentation for details.

To enable queries against the Global Catalog, simply configure the LDAP Server Information in the LDAP Authentication page to point to the IP address or host name of a Domain Controller that has the Global Catalog role enabled, and configure the LDAP port as 3268.

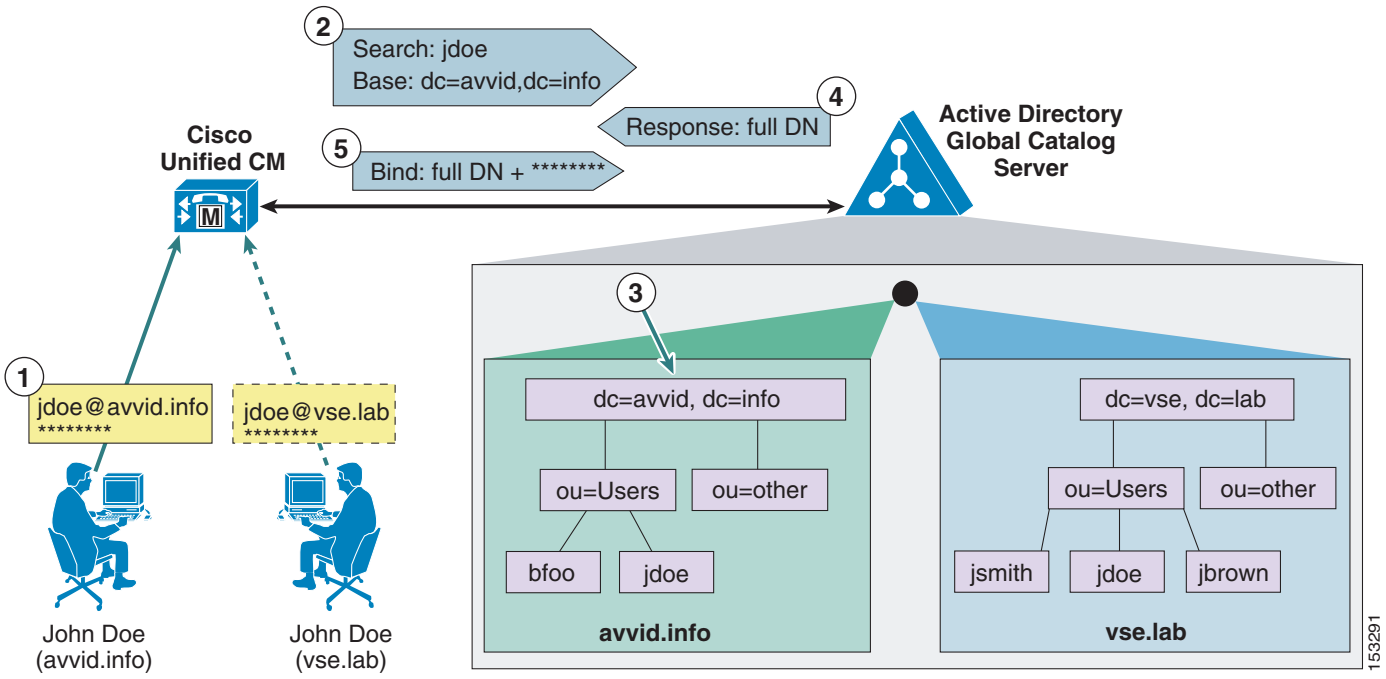
The use of Global Catalog for authentication becomes even more efficient if the users synchronized from Microsoft AD belong to multiple domains, because it allows Unified CM to authenticate users immediately without having to follow referrals. For these cases, point Unified CM to a Global Catalog server and set the LDAP User Search Base to the top of the root domain.

In the case of a Microsoft AD forest that encompasses multiple trees, some additional considerations apply. Because a single LDAP search base cannot cover multiple namespaces, Unified CM must use a different mechanism to authenticate users across these discontiguous namespaces.

As mentioned in the section on LDAP Synchronization, page 16-10, in order to support synchronization with an AD forest that has multiple trees, the UserPrincipalName (UPN) attribute should be used as the user ID within Unified CM. When the user ID is the UPN, the LDAP authentication configuration page within Unified CM Administration does not allow you to enter the LDAP Search Base field, but instead it displays the note, "LDAP user search base is formed using userid information."

In fact, the user search base is derived from the UPN suffix for each user, as shown in Figure 16-15. In this example, a Microsoft Active Directory forest consists of two trees, avvid.info and vse.lab. Because the same user name may appear in both trees, Unified CM has been configured to use the UPN to uniquely identify users in its database during the synchronization and authentication processes.

Figure 16-15 Authentication with Microsoft AD Forests with Multiple Trees



As shown in Figure 16-15, a user named John Doe exists in both the avvid.info tree and the vse.lab tree. The following steps illustrate the authentication process for the first user, whose UPN is jdoe@avvid.info:

1. The user authenticates to Unified CM via HTTPS with its user name (which corresponds to the UPN) and password.
2. Unified CM performs an LDAP query against a Microsoft Active Directory Global Catalog server, using the user name specified in the UPN (anything before the @ sign) and deriving the LDAP search base from the UPN suffix (anything after the @ sign). In this case, the user name is jdoe and the LDAP search base is "dc=avvid, dc=info".
3. Microsoft Active Directory identifies the correct Distinguished Name corresponding to the user name in the tree specified by the LDAP query. In this case, "cn=jdoe, ou=Users, dc=avvid, dc=info".
4. Microsoft Active Directory responds via LDAP to Unified CM with the full Distinguished Name for this user.
5. Unified CM attempts an LDAP bind with the Distinguished Name provided and the password initially entered by the user, and the authentication process then continues as in the standard case shown in Figure 16-14.

Note Support for LDAP authentication with Microsoft AD forests containing multiple trees relies exclusively on the approach described above. Therefore, support is limited to deployments where the UPN suffix of a user corresponds to the root domain of the tree where the user resides. AD allows the use of aliases, which allows a different UPN suffix. If the UPN suffix is disjointed from the actual namespace of the tree, it is not possible to authenticate Unified CM users against the entire Microsoft Active Directory forest. (It is, however, still possible to use a different attribute as user ID and limit the integration to a single tree within the forest.)

User Filtering for Directory Synchronization and Authentication

Unified CM provides an LDAP Query Filter to optimize directory synchronization performance. Cisco recommends importing those directory user accounts that will be assigned to Unified Communications resources. To allow for enterprise-wide UDS based service discovery, all users assigned to Unified Communications resources on any cluster in the enterprise need to be imported to all clusters in the enterprise. Differentiation between local and remote users is achieved by the **Home Cluster** setting on the Feature Group Template associated with the LDAP synchronization agreement that is used. When the number of directory user accounts exceeds the number supported for an individual cluster, filtering must be used to select the subset of users that will be associated on that cluster. The Unified CM synchronization feature is not meant to replace a large-scale corporate directory.

In many cases, a unique search base is all that is needed to control which accounts are synchronized. When a unique search base is not available, a custom LDAP filter might be required. The information in the following sections addresses both methods that can be used to optimize directory synchronization. When any mechanism is used to limit the accounts imported into Unified CM, the default directory lookup configuration will list only those directory entries that exist in the Unified CM database unless the UDS LDAP proxy functionality is used. For directory lookup to access the entire directory, you also can configure Unified CM to utilize an external web server. Details of this configuration are not discussed here but are discussed in the Unified CM product documentation available at

https://www.cisco.com/en/US/partner/products/sw/voicesw/ps556/tsd_products_support_series_home.html

Optimizing Unified CM Database Synchronization

The Unified CM Database Synchronization feature provides a mechanism for importing a subset of the user configuration data (attributes) from the LDAP directory store into the Unified CM publisher database. Once synchronization of a user account has occurred, the copy of each user's LDAP account information may then be associated to additional data required to enable specific Unified Communications features for that user. When authentication is also enabled, the user's credentials are used to bind to the LDAP store for password verification. The end user's password is never stored in the Unified CM database when enabled for synchronization and/or authentication.

User account information is cluster-specific. Each Unified CM publisher server maintains a unique list of those users receiving Unified Communications services from that cluster. Synchronization agreements are cluster-specific, and each publisher has its own unique copy of user account information. Only those users who will be assigned Unified Communications resources should be synchronized with Unified CM. The following is a partial list of common reasons why the entire set of users defined in the LDAP directory should not be imported into the Unified CM cluster:

- Importing users who will not be assigned Unified Communications resources can increase directory synchronization time.
- Importing users who will not be assigned Unified Communications resources can slow Unified CM searches and overall database performance.
- In many cases, the number of user accounts in the LDAP directory store far exceeds the total user capacity of the Unified CM database.

Unified CM has no enforced limit on the number of accounts that may be added to the system. Cisco recommends limiting the number of users to twice the supported number of endpoints. There might be cases where accounts are needed for applications, and some designs might require additional accounts.

Cisco recommends using the control mechanisms described here to minimize the number of user accounts imported, regardless of the LDAP database size. This will improve the speed of the first and subsequent periodic synchronizations and will also improve manageability of the user accounts.

Using the LDAP Structure to Control Synchronization

Many deployments of LDAP directories use the Organizational Unit Name (OU) to group users into a logical order and sometimes hierarchical order. If the LDAP directory has a structure that organizes users into multiple OUs, then it often is possible to use that structure to control the groups of users imported. Each individual Unified CM synchronization agreement specifies a single OU. All active accounts under the specified OU, even within sub-OUs, are imported. Only those users in the OU are synchronized. When multiple OUs containing users are required in a cluster, multiple synchronization agreements are required. When an OU contains users that will not be assigned Unified Communications resources, Cisco recommends omitting those OUs from the directory synchronization.

The same technique may be used with AD, which defines containers. A synchronization agreement may specify a particular container in the directory tree and thereby limit the extent of the import.

Because there is only a limited number of synchronization agreements available, LDAP deployments with many OUs or containers can quickly exhaust this technique. One possible method to synchronize users in a multi-OU environment is to control the permissions assigned to the synchronization service account. Configure the synchronization agreement to a tree node that contains a mix of users, and then restrict the system account from read access to selected parts of the subtree. Refer to your LDAP vendor documentation on how to restrict this access.

LDAP Query

Additional control over filtering might be required for any of the following reasons:

- The LDAP directory has a flat structure that does not enable adequate control by configuration of the synchronization agreements. When the aggregate number of users that are imported by all the synchronization agreements is greater than the maximum number of users supported by the Unified CM cluster, then it is necessary to control the number of users imported through filters.
- You want to import a subset of user accounts into the Unified CM cluster, for administrative segmentation of users, to control a subset of users that have access and authentication to the cluster. Any account that is imported into a cluster has some level of access to the web pages and authentication mechanisms, which might not be desirable in some cases.
- The LDAP directory structure does not have an accurate representation of how users are going to be mapped into the Unified CM clusters. For instance, if OUs are set up according to an organizational hierarchy but users are mapped to Unified CM by geography, there might be little overlap between the two.

In these cases, the LDAP Query filter may be used to provide additional control over the synchronization agreements.

LDAP Query Filter Syntax and Server-Side Filtering

Unified CM uses standard LDAP mechanisms for synchronizing data from an LDAP directory store. It utilizes the Search mechanism, as defined by RFC 4510 et seq., to send a request to retrieve data from the LDAP server. Also defined by that mechanism is the ability to specify a filter string inside the Search message that is used by the LDAP server to select entries in the database for which to return data. The syntax of the filter string is defined by RFC 4515 String Representation of Search Filters. This RFC may be used as a reference for constructing more complex filter strings.

The filter string is embedded within a Search message that is sent by Unified CM to the LDAP server and is executed by the server to select which user accounts will be provided in the response.

Simple Filter Syntax

You can configure a filter by specifying standard attribute names and values that are desired for those attributes. The attributes may also be specified by DN element instead of name. The filter string that is used by Unified CM in LDAP queries is stored internally in the ldapfilter table and is the string inserted into the Search message.

A filter is a UTF-8 formatted string that has the following syntax:

(*attribute operator value*)

or

(*operator*(*filter1*)(*filter2*))

Where *filter1* and *filter2* have the syntax shown in first line, and the *operator* is one of those listed in Table 16-6. The *attribute* corresponds to an LDAP attribute that exists in the directory, *operator* is one of the operators listed in Table 16-6, and *value* corresponds to the actual data value that is requested for the attribute.

Table 16-6 Basic Filter String Operators

Operator	Meaning of Function
!	Logical NOT
&	Logical AND
|	Logical OR
*	Wildcard
=	Equal to
>=	Lexicographically greater than or equal to
<=	Lexicographically less than or equal to

An attribute specified in the filter can be any attribute that exists in the LDAP directory store, and it does not have to be one of the attributes that is understood and imported by Unified CM. The attribute is used only on the LDAP server to select data, and the corresponding entries will have a subset of their data imported into Unified CM.

Example 16-1 A Single Condition

(givenName=Jack)

The filter in Example 16-1 selects any user with a given name of Jack.

Example 16-2 Multiple Conditions May Be Joined with Logical Characters

(&(objectclass=user)(department=Engineering))

The filter in Example 16-2 selects all users in the engineering department.

Default Filter Strings

If no custom filter strings are defined, Unified CM uses a default LDAP filter string as follows:

- Default Active Directory (AD) filter string

  (&(objectclass=user)(!(objectclass=Computer))(!(UserAccountControl:1.2.840.113556.1.4.803:=2)))

  This default filter selects entries for which the object class is a user but not a computer, and for which the account is not flagged as disabled.

- Default Active Directory Application Mode (ADAM) or Active Directory Lightweight Directory Services (AD LDS) filter string

  (&(objectclass=user)((objectclass=Computer))(!(msDS-UserAccountDisabled=TRUE)))

- Default filter string for all other directory types

  (objectclass=inetOrgPerson)

Extending the Default Filter

Cisco recommends that you use the default filter string and append additional conditions to it. For example:

(&(objectclass=user)(!(objectclass=Computer))(!(UserAccountControl:1.2.840.113556.1.4.803:=2))(telephonenumber=+1919*))

This filter selects only users that have a prefix of +1919 in their telephonenumber field. The synchronization agreement will import only users with an area code of 919 in the US. This example assumes all entries are in +E.164 format.

For the search filter, you may use any existing attribute or even a custom attribute that is defined in the LDAP directory store. The filter string controls which records are selected by the LDAP server to be returned to Unified CM, but the attributes that are imported are not affected by the filter string.

Custom LDAP filter strings can be up to 2048 characters long. Custom LDAP filters first need to be created, and then existing custom LDAP filters can be assigned to LDAP synchronization agreements. Different LDAP synchronization agreements can use different custom LDAP filters.

High Availability

Unified CM LDAP Synchronization allows for the configuration of up to three redundant LDAP servers for each directory synchronization agreement. Unified CM LDAP Authentication allows for the configuration of up to three redundant LDAP servers for a single authentication agreement. You should configure a minimum of two LDAP servers for redundancy. The LDAP servers can be configured with IP addresses instead of host names to eliminate dependencies on Domain Name System (DNS) availability.

Capacity Planning for Unified CM Database Synchronization

The Unified CM Database Synchronization feature provides a mechanism for importing a subset of the user configuration data (attributes) from the LDAP store into the Unified CM publisher database. Once synchronization of a user account has occurred, the copy of each user's LDAP account information may then be associated to additional data required to enable specific Unified Communications features for that user. When authentication is also enabled, the user's credentials are used to bind to the LDAP store for password verification. The end user's password is never stored in the Unified CM database when enabled for synchronization and/or authentication.

User account information is cluster-specific. Each Unified CM publisher server maintains a unique list of those users receiving Unified Communications services from that cluster. Synchronization agreements are cluster-specific, and each publisher has its own unique copy of user account information.

The maximum number of users that a Unified CM cluster can handle is limited by the maximum size of the internal configuration database that gets replicated between the cluster members. The maximum number of users that can be configured or synchronized is 160,000. With more than 80,000 users the maximum number of LDAP synchronization agreements is limited to 10, while with less than 80,000 users the total number of LDAP synchronization agreements is limited to 20. To optimize directory synchronization performance, Cisco recommends considering the following points:

- Directory lookup from phones and web pages may use the Unified CM database, the IP Phone Service SDK, or the UDS LDAP proxy functionality. When directory lookup functionality uses the Unified CM database, only users who were configured or synchronized from the LDAP store are shown in the directory. If a subset of users are synchronized, then only that subset of users are seen on directory lookup.
- When the IP Phone Services SDK is used for directory lookup, but authentication of Unified CM users to LDAP is needed, the synchronization can be limited to the subset of users who would log in to the Unified CM cluster.
- If only one cluster exists, and the LDAP store contains fewer than the maximum number of users supported by the Unified CM cluster, and directory lookup is implemented to the Unified CM database, then it is possible to import the entire LDAP directory.
- When multiple clusters exist and the number of users in LDAP is less than the maximum number of users supported by the Unified CM cluster, it is possible to import all users into every cluster to ensure directory lookup has all entries.
- If the number of user accounts in LDAP exceeds the maximum number of users supported by the Unified CM cluster and the entire user set should be visible to all users, it will be necessary to use the Unified IP Phone Services SDK to off-load the directory lookup from Unified CM.
- If both synchronization and authentication are enabled, user accounts that have either been configured or synchronized into the Unified CM database will be able to log in to that cluster. The decision about which users to synchronize will impact the decision on directory lookup support.

Note Cisco supports the synchronization of user accounts up to the limit mentioned above, but it does not enforce this limit. Synchronizing more user accounts can lead to starvation of disk space, slower database performance, and longer upgrade times.

UDS Proxy for LDAP

Clients that use User Data Service (UDS) for contact source access are limited to accessing only the users that exist in the Unified CM end-user database. Although this database can be populated from the corporate directory via LDAP synchronization, the number of users searched is still limited by the maximum number of users supported in Unified CM. (See the section on Capacity Planning for Unified CM Database Synchronization, page 16-31, for details.) To overcome this limitation, Cisco Unified CM 11.5 and later releases can be set up to act as a UDS-to-LDAP proxy for UDS-based user searches. In this mode, for every user search requested via UDS, Unified CM connects back to the corporate directory to execute the search and then relays the results back to the client via UDS. Instead of serving the UDS search requests directly, Unified CM in this mode relies on the information returned from the corporate LDAP directory.

Note The recommended contact source for on-premises Jabber deployments is Cisco Directory Integration (CDI). UDS proxy for LDAP should be used only for deployments where endpoints rely on UDS for contact searches and where the number of users in the corporate directory exceeds the maximum number of end users supported in Cisco Unified CM.

UDS proxy functionality is enabled globally in Unified CM. Up to three directory Unified Communications services can be selected to define the LDAP data sources for the proxy functionality. Unified CM uses an LDAP bind to execute the search operations, and the user and password to be used for this bind operation are configured specifically for this feature. The UDS proxy supports up to three LDAP user search bases.

Directory Integration for VCS Registered Endpoints

Cisco TelePresence Video Communication Server (VCS) endpoints are managed by VCS and as such they can receive directory information from the Cisco TelePresence Management Suite (TMS). Cisco TelePresence Management Suite offers many more services such as scheduling of Unified CM and VCS registered endpoints, and management of VCS registered endpoints.

Cisco TelePresence Management Suite can manage multiple phone books coming from multiple sources.

Cisco TMS 14.1 can also integrate with Cisco Unified Communications Manager and receive directory information from Unified CM. This is the recommended configuration in order to have a unified directory for Unified CM and VCS endpoints.

Multiple Unified CM clusters can be added as multiple directory sources to Cisco TMS and organized in a single directory. TMS can push directory information to endpoints connected to it and registered to a single VCS or to multiple VCSs.

For more information, refer to the latest versions of the *Cisco TelePresence Management Suite Administrator Guide* and the *Cisco TelePresence Management Suite Provisioning Extension Deployment Guide*, both available at

https://www.cisco.com/en/US/products/ps11338/tsd_products_support_series_home.html

Identity Management Architecture Overview

Figure 16-16 presents an overview of the identity management architecture. All Cisco Collaboration Applications (for example, Cisco Unified CM with IM and Presence, and Cisco Unity Connection) maintain their individual identity stores. Users in these identity stores can be synchronized from the enterprise directory by means of individual LDAP sync agreements, but they can also be configured locally. Synchronizing from LDAP is highly recommended to make sure that all relevant principals (users) exist both in the corporate directory and in the individual identity stores.

LDAP synchronization is a prerequisite to be able to use single sign-on (SSO) for collaboration clients and workstations accessing administration interfaces or the various Unified Communications services provided by the collaboration applications. SSO is implemented based on Security Assertion Markup Language (SAML) version 2.0 (SAML 2.0). SAML 2.0 authentication uses SAML authentication flows between the clients accessing the services, the collaboration applications providing these services, and an Identity Provider (IdP). The IdP is the component responsible for the actual authentication of users. The IdP can support various authentication mechanisms, including user/password based authentication against LDAP, Kerberos authentication, SmartCard based authentication, and others. The IdP can be any

IdP available on the market that complies with the SAML 2.0 specification. Cisco validates SSO with some of the IdPs such as OpenAM, Ping Federate, and Microsoft Active Directory Federated Services (ADFS).

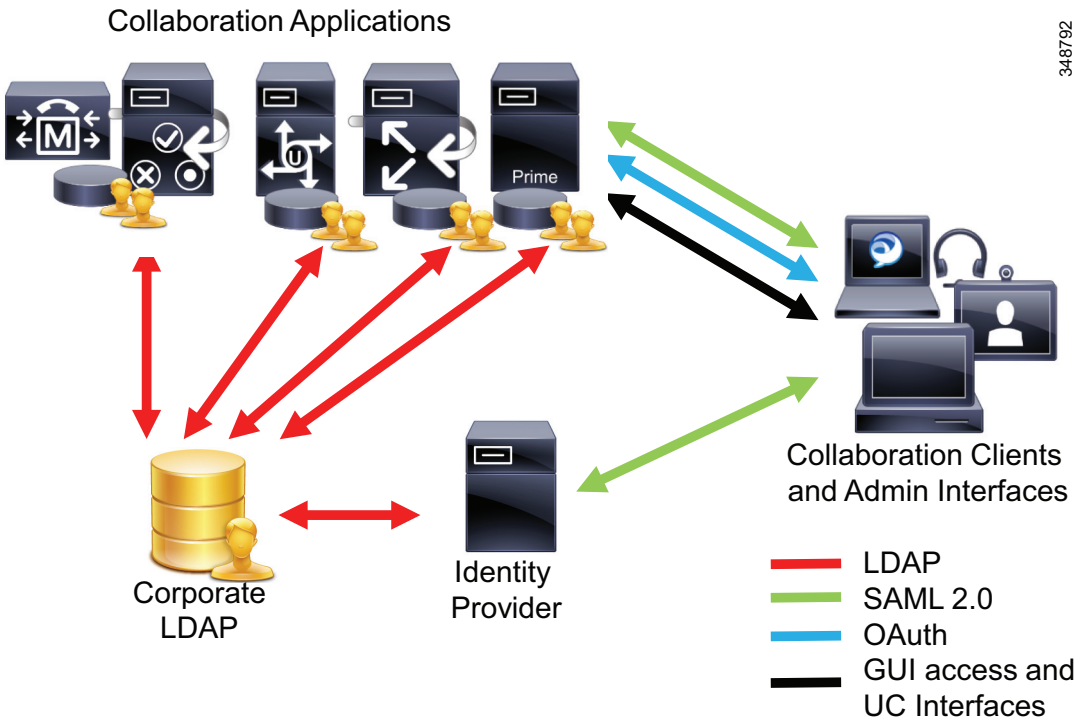
For single sign-on (SSO), authentication to Unified Communications services is delegated to the IdP through SAML 2.0. Using this mechanism, any user has to authenticate only once to any of the entities providing Unified Communications service and can then access all other Unified Communications service providers' GUIs without having to authenticate again.

SAML 2.0 only provides authentication of users and is browser based. SAML 2.0 does not address requirements for distributed authorization to use any of the UC interfaces including Unified CM UDS, Unified CM SIP, Unified CM CTI, Unified CM IM and Presence SOAP, Unified CM IM and Presence XMPP, and Unity Connection VMRest.

A centralized authorization service running on Unified CM provides authorization for UC services offered by Unified CM, Unified CM IM and Presence, and Unity Connection. This centralized authorization functionality is based on the OAuth 2.0 specification. OAuth 2.0 is an open framework for authorization providing delegated access to services on behalf of resource owners. The protocol enables authorization of clients to access resources without sharing the credentials used by the clients for authentication. OAuth essentially allows access tokens to be issued to clients by a central authorization instance. These tokens then are presented by the clients to the servers offering a service, as proof of authorization. An access token is a string value that represents the granted level and duration of access to specific resources. The access token either only represents an identifier that can be used to retrieve details of the authorization from the authorization service, or it may contain the actual details of the authorization. In the latter case, the access tokens are called *self-contained*, and the content of the access token must be signed so that the authenticity of the authorization details can be verified. Optionally, the content of self-contained tokens can also be encrypted. If the access tokens are not self-contained, then the service providers need to ask the authorization service for validation of the presented tokens to verify the authorization. In this process the content of the authorization token, as well as the mechanism used to authenticate the client and issue the token, is completely transparent for the service provider.

The authorization framework of the Cisco Collaboration solution uses self-contained access tokens. The keys used to sign and encrypt the access tokens are pushed from Unified CM to Unified CM IM and Presence, and they are pulled from Unified CM by Cisco Unity Connection and Cisco Expressway.

Figure 16-16 Identity Management Architecture



Single Sign-On (SSO)

SSO via SAML 2.0 is an additional authentication option to the previously existing LDAP bind and local authentication. For SAML 2.0 SSO, all Unified Communications services, including the OAuth authorization service, integrate directly with the corporate identity management system using SAML 2.0.

The primary protocol used for SSO is SAML. Detailed information about SAML, such as protocol specification, use cases, and authentication flows, is openly available on the Internet. This section only introduces some key aspects of SAML.

All interactions with an Identity Provider (IdP) using SAML must be through a Web browser on the client side. If SAML authentication is to be used for clients that do not expose a Web GUI to the user, then these clients use internal WebView clients. Examples of this include Jabber softclients and collaboration endpoints supporting SSO.

Security Assertion Markup Language (SAML) is an XML data format specifically designed for the data exchange between service providers (SPs) and an IdP. SAML uses security tokens containing assertions to pass authentication related information between the IdP and the SP. The IdP in this exchange takes the role of a SAML authority, whereas the SP is a SAML consumer. Specifications of SAML can be found at

https://saml.xml.org/saml-specifications

Before SAML authentication can take place, a trust relationship between the service providers (SPs) and the Identity Provider (IdP) has to be established. This is done by exchanging metadata between the SP and IdP.

In general, a single SAML metadata instance describes either a single SAML entity or multiple entities. A SAML metadata instance describing multiply SAML instances contains a list of descriptions of single entities. Prior to Cisco Unified CM release 11.5, SAML metadata instances created by Cisco Collaboration solutions always describe only a single SAML instance.

For any SAML instance described by a SAML metadata instance, the metadata contains:

- A unique identifier
- Organization
- Expiration time for this information
- Caching period
- XML signature of this information
- Contact persons
- Unique identifier of the entity (entity ID)
- Description of SAML role of this SAML instance (identity provider, service provider, and so forth)

All pieces except the unique identifier are optional in the SAML specification and are not included in metadata created by Cisco collaboration SPs.

Each role description included in a SAML metadata instance defines the supported protocols and optionally also contains SSO key information. These keys are used later to sign SAML messages exchanged between SAML entities.

SAML metadata for a SAML service provider is required by SAML identity providers to understand the aspect of the service provider relevant for the SAML exchange between these two entities. The portion of SAML metadata specific to the service provider can indicate whether the service provider will sign SAML authentication requests and whether the service provider expects SAML assertions returned to the service provider to be signed. Also, the service provider SAML metadata defines where the authentication response should be posted. This authentication consumer service (ACS) definition basically is a URL. In addition, the service provider SAML metadata might define attributes to be exchanged between the SAML service provider and the identity provider as part of the SAML authentication process.

Similarly, identity provider metadata defines the IdP characteristics relevant for the SAML exchange between IdP and SP. IdP metadata also can define signing requirements for authentication requests and what attributes should be exchanged between IdP and SP as part of the SAML authentication process.

Detailed information about the SAML metadata format can be found at

https://saml.xml.org/saml-specifications

SAML metadata created by Cisco Collaboration SPs contains only:

- ID, entityID: Both set to the FQDN of the node or, in case of cluster-wide SSO, the FQDN of the publisher node.
- AuthnRequestSigned: **false**. This indicates that authentication requests sent by this entity will not be signed unless requested otherwise by the IdP.
- WantAssertionsSigned: **false**. This indicates that SAML assertions do not need to be signed to be accepted by this entity, but signed assertions are also acceptable.
- Encryption key and signing key: The metadata contains the node's Tomcat certificate for both keys. In the case of cluster-wide SSO, the multi-server Tomcat certificate is used, which requires the cluster to be configured to use a multi-server Tomcat certificate.

- nameIDFormat: **transient**. This indicates that name identifiers used to identify subjects in SAML assertions will be transient, which means that these identifiers cannot be used to identify a subject because the IdP will issue a new unique opaque identified the next time the same subject authenticates successfully. Instead, the authenticated subject will be identified based on the **uid** attribute returned by the IdP.
- AssertionConsumerService: One or multiple assertion consumer service definitions are included. Prior to Cisco Unified CM release 11.5, only a singe assertion consumer service definition for the HTTP-POST binding is included. Starting with Unified CM release 11.5, one assertion consumer service definition for each node and binding (HTTP-POST and HTTP-Redirect) is included. Each assertion consumer service definition specifies the binding (HTTP-POST or HTTP-Redirect) and the URL of the assertion consumer service (for example: https://ucm.example.org:8443/ssosp/saml/SSO/alias/ucm.example.org).

SAML Authentication

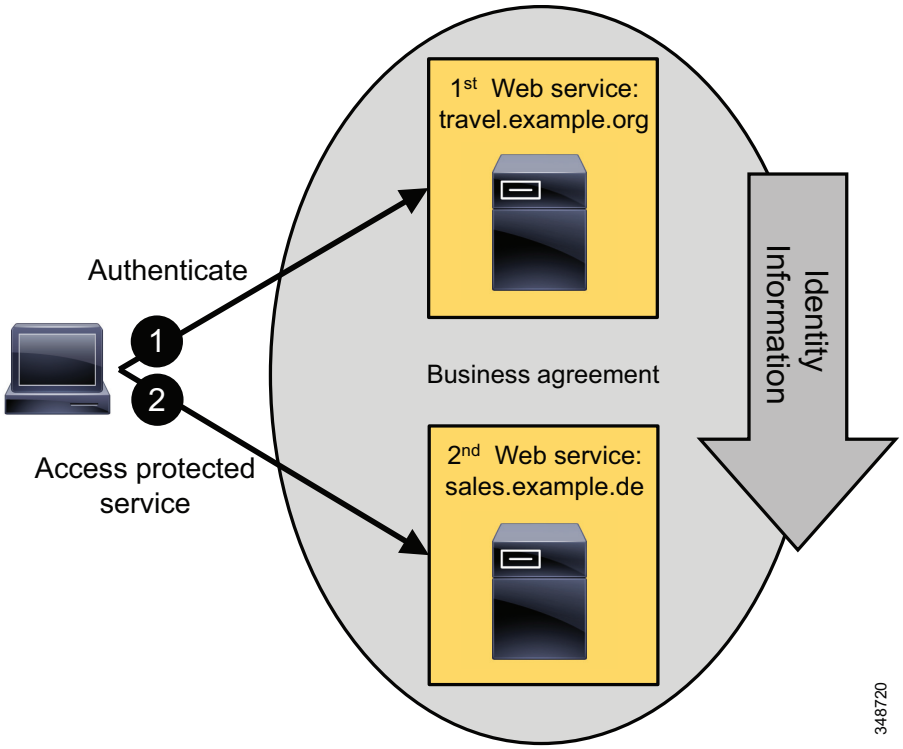
The actors in generic SAML authentication flow are:

- Client — Browser-based user client used to access the service
- SP — Application or service the user tries to access
- IdP — Entity performing the user authentication based on user credentials. The actual credentials and the actual authentication mechanism are hidden by the IdP. The IdP issues SAML assertions based on the authentication process result.

SAML defines a number of profiles to describe the use of SAML to solve typical use cases. The relevant profile used for SSO with Cisco Collaboration services is the web browser SSO profile of SAML V2.0.

The use case solved by this profile is the multi-domain web single sign-on, illustrated in Figure 16-17. In this use case, a user already has a login session with some web service (for example, travel.example.org) and is using this service. As part of the login process, a security context has been established for travel.example.org. If the same user now moves to another web service (for example, sales.example.de) and a business agreement exists between travel.example.org and sales.example.de that establishes a federated identity for the user between these services, then the user is able to access the web service sales.example.de without having to provide authentication credentials again. In this case the identity provider site (travel.example.org) asserts to the service provider site (sales.example.de) that the user is known, has been properly authenticated, and has certain identity attributes. The service provider site (sales.example.de) trusts this assertion based on the existing business agreement between the sites and grants access to the service.

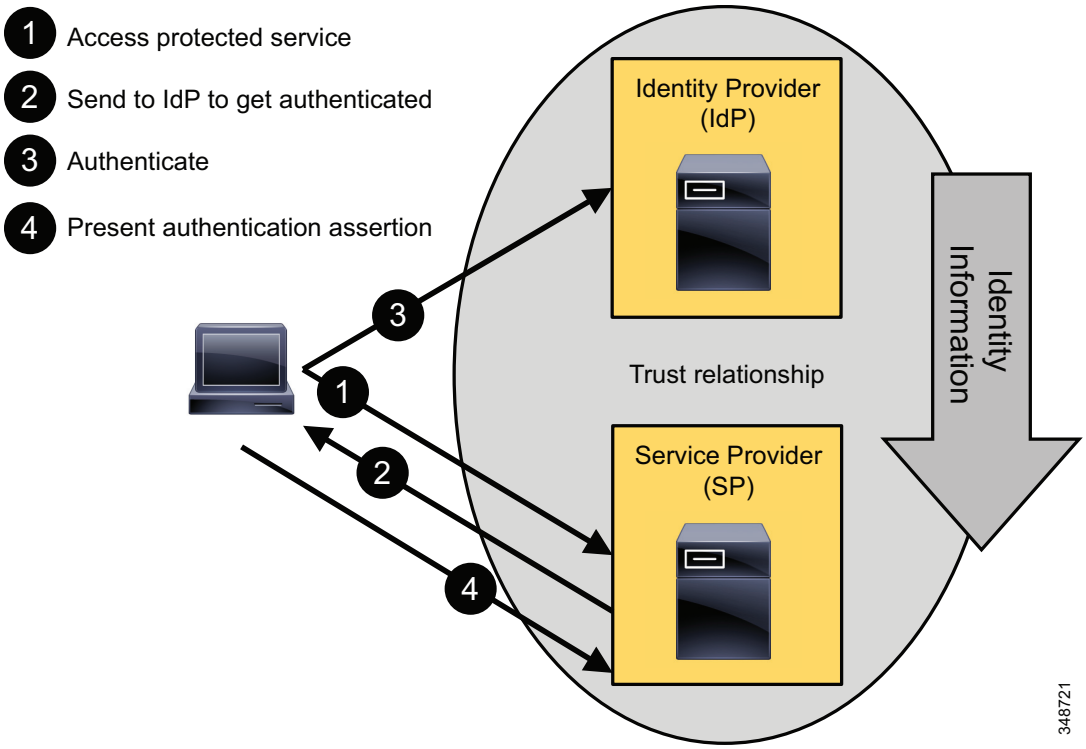
Figure 16-17 Multi-Domain Web Single Sign-On



This description implies that the user first is authenticated by a web service and that this first web service then provides an identity assertion to enable the user to access the second web service. The web service accessed first (travel.example.org) acts as the IdP for SP sales.example.de. This is known as IdP initiated web SSO.

The more typical web SSO flow used with Cisco Collaboration Services is SP initiated web SSO, illustrated in Figure 16-18. In this case the user directly (without visiting an IdP first) tries to access a protected resource on an SP. The SP sends the user to the IdP to get authenticated, and then the user presents the authentication assertion received from the IdP to the SP to get access.

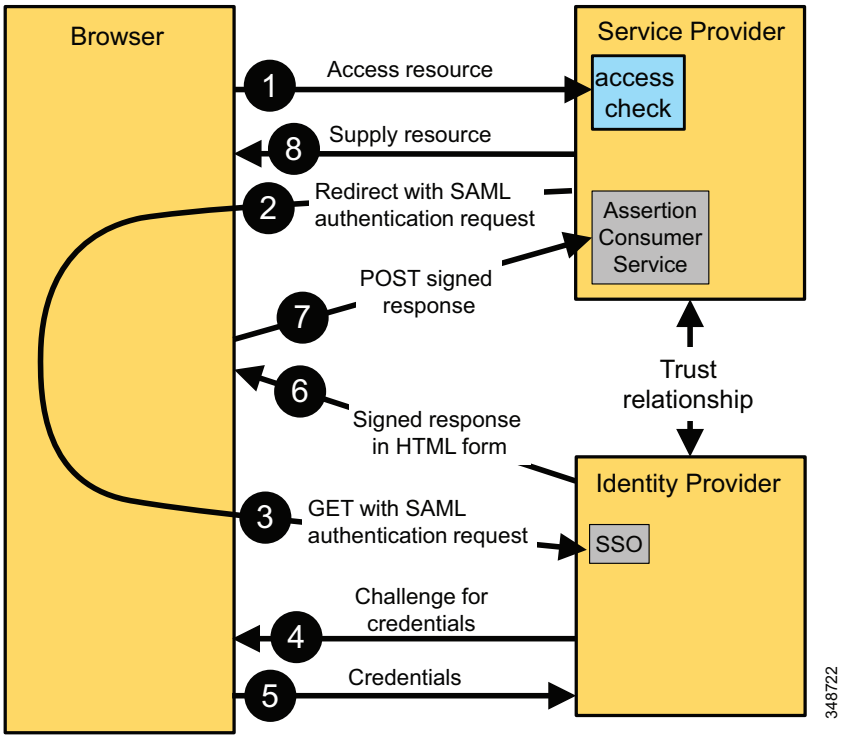
Figure 16-18 Web SSO Initiated by a Service Provider



The SAML web browser SSO profile provides a variety of options depending on whether the authentication is initiated by the IdP or SP and on how the messages are exchanged between IdP and SP. As mentioned above, Cisco Collaboration services use SP initiated SSO only where the SP sends a user to an IdP first to authenticate when the user is trying to access a protected resource and does not have an active session with the service provider. The IdP then builds an authentication assertion and sends the user back to the SP with that assertion.

The binding used for the messages exchange between IdP and SP for Cisco Collaboration services is the Redirect/POST binding, illustrated in Figure 16-19. Here an HTTP 302 redirect is used to send the SAML authentication request message from the SP to the IdP, and the authentication response from IdP to SP is sent using an HTTP POST message.

Figure 16-19 SP-Initiated SSO (Redirect/POST Binding)



The general steps of the SAML authentication flow are:

1. The user tries to access a service or resource by pointing the browser to the URL hosted on the application server. The browser at this moment does not have an active session with the service.
2. The SP realizes that the request originates from a client without an active session. Because HTTP is stateless, an active session can be detected by the SP only if the client sends a session cookie that has been issued by the SP earlier. Based on the SSO configuration, the SP now generates an SAML authentication request to be sent to the appropriate IdP defined as part of the SSO configuration. The SAML request contains information about the SP generating the request. This is required so that the IdP can identify the SPs sending SAML requests.

   The SP does not communicate directly with the IdP to authenticate the user. Instead the SP redirects the browser to the IdP. The URL used for this redirect is taken from the IdP metadata exchanged earlier. The SAML request to be sent to the IdP is included in the redirect as a URL query parameter using Base64 encoding.

   This redirecting HTTP 302 might look like this:

```
HTTP/1.1 302 Found
Location:
https://pingsso.example.com:9031/idp/SSO.saml2?SAMLRequest=nZLNbtswEITveQqCd1m0pKo
WY
RlwYxQ1kDZK5OaQG02tYwISqXLJtH37kkra%2FBjwodflcPab3V2iGPqRr7076lv44QEdIb%2BGXiOfXm
rqreZGoEKuxQDIneTt%2BusVz2aMj9Y4I01PL7abmmJWVCxnku07sYCqFAu2KGWVdaycV1AWRbnPPjJZl
Dkld2BRGV3TYEPJFtHDVqMT2oUSm%2BcJq5Ks2LGK5x84K%2B8p2QQ0pYWbfh2dG5Gn6aj0A6KZHc0AM2
MfeACYp6ob07a9nsUEGSWfjZUwJazpQfQIsWEjENUj%2FKs0z1E%2BKd0F0%2FO5908i5F92uyZprtsdJ
WtEsJHu0mj0A9gW7KOS8P326oVXejkk4F94F0WRpyEBjmmkjdip6JXAEyldXSyjhE%2FDsq%2BWdJ5V%2
FOWiq%2FeWy%2FSV4bP9yL8Fi%2B2mMb2Sv%2F%2FnFuK8B%2BHOq2NFdclhknJnhUYF2lHSNrH%2FjQ9
DOCiwNT2ZA1n3vfl5aUG4sD5nPdDVU5K37CFQenrdqz8%3D&RelayState=s249030c0bda8e96a8086c
92d0619e6446b270c463
```

The encoded SAML authentication request shown above can be decoded to:

```
<samlp:AuthnRequest xmlns:samlp="urn:oasis:names:tc:SAML:2.0:protocol"
    ID="s249030c0bda8e96a8086c92d0619e6446b270c463"
    Version="2.0"
    IssueInstant="2013-09-19T09:35:06Z"
    Destination="https://pingsso.example.com:9031/idp/SSO.saml2"
    ForceAuthn="false"
    IsPassive="false"
    ProtocolBinding="urn:oasis:names:tc:SAML:2.0:bindings:HTTP-POST"
    AssertionConsumerServiceURL="https://cucm-eu.example.com:
8443/ssosp/saml/SSO/alias/cucm-eu.example.com"
    >
<saml:Issuer xmlns:saml="urn:oasis:names:tc:SAML:2.0:assertion">
cucm-eu.example.com</saml:Issuer>
<samlp:NameIDPolicy xmlns:samlp="urn:oasis:names:tc:SAML:2.0:protocol"
    Format="urn:oasis:names:tc:SAML:2.0:nameid-format:transient"
    SPNameQualifier="cucm-eu.example.com"
    AllowCreate="true"
     />
</samlp:AuthnRequest>
```

Among other details specifying authentication parameters and identifying the requesting SP, the above SAML authentication request also specifies the Assertion Consumer Service (ACS) URL. The ACS URL is the URL to which the SAML Authentication Response needs to be POSTed at the end of the authentication process.

3. The browser receives the redirect, follows the URL, and issues the corresponding GET to the IdP. The SAML request is maintained. The browser at this stage does not have an active session with the IdP.

4. After receiving the new request from a browser with no active session (browser is not sending a cookie issued by the IdP earlier), the IdP authenticates the user based on the pre-configured authentication mechanisms. Possible authentication mechanisms include user/password, PKI/CAC, or Kerberos. For user/password authentication, the IdP might push a form to the user to enter the credentials (for example, “200 OK” message with an IdP login form). For the actual authentication, the IdP might depend on back-end systems such as an LDAP server for user/password authentication.

   One key point here is that the exchange of credentials for the purpose of authentication takes place between the IdP and the browser. The SP is not involved and does not see the credentials.

5. The browser provides further information required for the authentication process. For the user/password case, this would be a POST with the information. For other authentication mechanisms, other details would need to be sent to the IdP by the browser.

6. The IdP now checks and validates the provided credentials. The check could involve interactions with respective back-end systems (LDAP bind for user/password authentication against LDAP, communication with Kerberos server to validate ticket, and so forth).

   Finally the IdP generates an SAML response for the SP. This response contains the SAML assertion documenting the result of the authentication process. The SAML assertion, in addition to the basic Yes/No information, also contains validity information and information about attributes describing the authenticated entity. At least the user ID of the authenticated entity has to be included in the well known attribute **uid** so that the SP can extract this information from the assertion to relate the authenticated entity to users existing in the local database.

   The SAML assertion is signed and potentially encrypted by the IdP according to the SSO key information published in the IdP metadata. This ensures that the SP can verify the authenticity of the SAML assertion.

The IdP returns the SAML assertion to the browser in a hidden form in a 200 OK message. The hidden form instructs the browser to POST the SAML assertion to the Assertion Consumer Service (ACS) URL of the SP.

The IdP has to establish a security context so that future authentication requests from the same browser can be answered without going through the exchange of credentials. The IdP will then realize that it already has a valid session with the browser and will assert the authentication of the previously authenticated user without prompting for credentials again. This context is established via a session cookie set on the browser by the IdP. This basically enables SSO for multiple SPs.

7. The browser follows the hidden POST received in the 200 OK message and POSTs the SAML assertion to the Assertion Consumer Service on the SP.
8. The SP extracts the SAML assertion from the POST and validates the signature of the assertion. This guarantees the authenticity of the SAML assertion and the IdP. The user identifier received in the SAML assertion in attribute **uid** as part of the attribute statement is then used to decide whether the user is authorized to access the requested service. This is based on local access control configuration on the SP. The **uid** value received in the SAML assertion has to match the Unified CM user ID of an end user authorized for the requested service. To make sure that user identifiers sent by the IdP in SAML assertions correlate to user IDs in Unified CM, SSO authentication is supported only for end users synchronized from LDAP. The assumption here is that the IdP is integrated with the same directory, so that **uid** values returned by the IdP are based on the same data source as Unified CM end user information.

   The SP grants access to the requested resource and sends back the content in a 200 OK message to the browser. The SP also sets a session cookie in the browser so that, for subsequent access requests from the same browser to the same SP, the SP does not have to initiate any more exchanges with the IdP. The IdP will be involved with additional requests from the same browser only after the SP session has expired.

Authentication Mechanisms for Web-Based Applications

When SSO is enabled for a collaboration service, any access to the respective service will be authenticated using SSO. As a fallback measure, a vanity or recovery URL also exists on the landing page. The vanity URL bypasses the SSO mechanism and provides access to all administrator GUIs. Access to the administrator GUI through the vanity URL is authenticated against the local user database. Access to the GUI through the vanity URL can be disabled on the CLI using the **utils sso recovery-url disable** command.

The vanity URL can be used as a recovery back door when there is an issue with the SAML infrastructure, such as when the IdP is unreachable or down, when there are metadata issues (for example, expired signing certificates), or when there are IdP configuration changes.

Collaboration services currently support the following user types:

- OS user

  This user is specified during installation and has access to the CLI, the Disaster Recovery System (DRS) GUI, and the OS Admin GUI. Credentials for OS users are maintained separately from credentials of other users. When enabling SSO, access to the CLI always is authenticated locally using the password stored in the local database, while access to the DRS and OS Admin GUI is authenticated via SSO and authorized against the platform database. The **set account name** command on the CLI can be used to create mappings from SSO UID values to platform users.

- Application user

  These are functional users created and managed locally. Passwords are stored in the local database. Application users are not enabled for SSO. With SSO enabled, application users can get access to only the Admin GUI through the vanity URL on the landing page.

- Local end user

  These users are created and managed locally. Passwords are stored locally. These users do not exist in the enterprise identity management system. If SSO is enabled, local end users cannot authenticate successfully. Local end users and LDAP synced users without SSO enabled are still supported.

- LDAP synced end user

  These users are managed in the corporate LDAP directory and are synchronized into the Unified Communications service through LDAP sync agreements. For every LDAP synced end user in the local database, there is a matching user in the corporate LDAP directory. If SSO is disabled, the passwords of LDAP synced end users are validated through an LDAP bind operation. With SSO enabled, the authentication of LDAP synced users is based on the authentication mechanism defined on the IdP, and authorization is based on local configuration. An LDAP synced end user has to have the proper rights assigned locally to be able to access the requested resource.

PIN-based authentication is always (even with SSO enabled) based on local configuration. Multiple collaboration services maintain individual PINs. Starting with Cisco Unified CM release 11.5, PINs can be synchronized between Unified CM and Cisco Unity Connection.

The following web services are enabled for SSO based on SAML IdP redirects:

- Cisco Unified Communications Manager Admin GUI
- Cisco Unified CM Self Care Portal
- Cisco Unified Communications Manager Serviceability GUI
- Cisco Unified Communications Manager Reporting Tool GUI
- Cisco Unified Communications Manager Platform Admin GUI
- Cisco Unified Communications Manager Disaster Recovery GUI
- Cisco Unified Communications Manager IM and Presence Admin GUI
- Cisco Unified Communications Manager IM and Presence Platform Admin GUI
- Cisco Unity Connection Admin GUI
- Cisco Unity Connection Platform Admin GUI
- Cisco Unified Personal Communicator Assistant
- Cisco Unity Connection WebInBox

SSO for Cisco Jabber

All Jabber platforms use embedded browser controls for SSO. These controls rely on underlying operating system browser technologies (see Table 16-7). Even if the embedded browser controls used by Jabber are based on the same technology as the system browser, cookies are never shared between the system browser and the embedded controls used by Jabber. This implies that Jabber always requires dedicated authentication via SSO even if the user already authenticated against the same IdP using the system browser. The only way around this is to not set up the IdP to use persistent cookies instead of session cookies. This is not considered to be the best practice because by using persistent cookies the IdP authentication state gets exposed openly and can potentially be hijacked by other applications with access to the same cookie jar.

Table 16-7 Browser Technology and Cookie Sharing

OS	Windows	Mac OS	Apple iOS	Android
Underlying Browser	Internet Explorer	Safari	WebKit or Safari	WebKit
Control shares cookies with native OS browser	No	No	No	No

For Jabber on Apple iOS, the browser selection (WebKit or Safari) is determined by the **SSO Login Behavior for iOS** enterprise parameter. Choosing Safari as the browser for SSO on Apple iOS allows access to the iOS certificate store and other protected resources to which only Apple applications have access. This selection is required if certificate-based authentication schemes should be used via SSO.

Design Considerations for SSO

SAML SSO always has to be enabled or disabled for all nodes in a cluster. Either all nodes or no nodes in a cluster are enabled for SSO. Enabling SSO through the Admin GUI automatically enables SSO for all existing nodes at the same time. As part of this process, SP metadata is downloaded to be used to establish the circle of trust between the IdP and the cluster node(s).

Prior to Cisco Unified CM release 11.5, each cluster node had to be represented on the IdP as an individual SP. If a node was added later to a cluster that already was in SSO mode, the metadata of that added node had to be imported into the IdP to complete the list of defined SPs on the IdP.

Starting with Cisco Unified CM release 11.5, SSO can be enabled in cluster-wide mode. In this case only a single metadata file needs to be exchanged with the IdP. Cluster-wide SSO can be used only if a single multi-server Tomcat certificate is used on the cluster because the SAML metadata of the cluster can contain only a single encryption and signing key. Multi-server certificates need to be CA signed certificates. When new nodes are added to a cluster enabled for cluster-wide SSO, updated metadata has to be exchanged with the IdP to make sure that the IdP is aware of the new assertion consumer service URLs of the added nodes.

SP metadata of Cisco Collaboration SPs contains assertion consumer service definitions for HTTP-Post and HTTP-Redirect bindings. These bindings must be supported by and enabled on the IdP. For SSO in cluster-wide mode, the IdP must be able to support multiple assertion consumer service definitions for a single SP.

IdP metadata has to be imported into all SAML SPs. When SSO is enabled on the Admin GUI, the IdP metadata provided in the process is automatically imported on all nodes of the cluster. If assertion signing or encryption is used by the IdP, then the signing and encryption keys must be included in the IdP metadata exchanged with the Cisco Collaboration SPs.

The SP metadata does not include the optional ContactPerson information; therefore, the IdP will not be able to expose contact information for Cisco Collaboration SPs.

SAML SP can request signed assertions from the IdP by including WantAssertionsSigned in the SAML AuthnRequest. Currently Cisco Collaboration SPs do not send this information, and the same parameter is set to **False** in the SP metadata. This gives the IdP full control over assertion signing. Cisco recommends activating SAML Assertion signing on the IdP.

If not requested otherwise by the IdP, Cisco Collaboration SPs do not encrypt or sign SAML authentication requests. This must be supported by the IdP.

Cisco Collaboration SPs request namedid-format:transient in both the metadata and the SAML authentication requests. IdPs must support this format and must be configured accordingly.

As part of a SAML assertion, the IdP in the AttributeStatement must return an attribute **uid**, and the value of this attribute must match the user ID of the respective end user in Unified CM.

Availability of the IdP is a key requirement when using SSO. The IdP has to be deployed with full redundancy and fault tolerance. Essential for this kind of deployment is that the IdP is deployed with a single logical URL and that suitable load balancers and web server farms are deployed to make sure that the single IdP URL is highly available. The single IdP URL is included in the IdP metadata and is imported into all Cisco Collaboration SPs. Failure of a single element (for example, a single web server) should be invisible to the Collaboration service.

SAML requests and assertions are signed using SP and IdP certificates. The lifetime of these certificates has to be closely monitored to make sure that the SAML SSO mechanism continues to work.

SAML assertions contain validity information (NotBefore, NotOnOrAfter). To make sure that valid assertions are not rejected due to timing issues, it is essential to synchronize all services using appropriate mechanisms such as Network Time Protocol (NTP).

Authorization Framework

For users accessing web interfaces, authentication is either based on local configuration, based on LDAP or, in the case of SSO, based on the SAML exchange between the user's browser, the web server, and the IdP. After successful authentication, the web server (for accessing the Unified CM administration GUI this would be the administration application running on Unified CM) consults the local configuration to determine whether the authenticated user is authorized to access the given resource. For example, if user Bob when authenticating via SSO provided valid credentials to the IdP and thus authenticated successfully, then Unified CM could still deny access to the Unified CM administration interface if Bob is not a member of the "Standard CCM Super Users" group. In this case Bob would get only a prompt indicating that he does not have the required privileges to access the system, instead of getting access to Unified CM administration. Authorization for access to web services such as the Unified CM administration GUI or end-user pages always is based on access levels defined on the application.

Jabber clients and other endpoints require access to a number of collaboration interfaces (for example, Unified CM SIP, Unified CM CTI, Unified CM IM and Presence SOAP, and Unity Connection VMRest). To avoid multiple authentication mechanisms (per interface), the Cisco Collaboration system uses the OAuth authorization framework for centralized authorization based on a single authentication.

OAuth 2.0

The OAuth 2.0 authorization framework is an open standard defined by the IETF OAuth working group, and the current version of the standard has been released as RFC 6749.

Whenever applications need to access multiple services on behalf of users without OAuth, separate authentications and authorizations per service are required. This creates a suboptimal user experience for end users because they have to manage multiple sets of credentials (can be partially addressed by using SSO), and it also creates a trust problem in that an end user has to share the access credentials with multiple applications.

OAuth addresses these challenges by enabling applications to obtain access to a service on behalf of an end user by orchestrating an approval interaction between the end user and the service. As part of this approval interaction the end-user, after being authenticated, instructs the OAuth authorization service to grant access tokens to the application asking for authorization. The application then presents an access token as proof of authorization when accessing services.

Access tokens have a limited lifetime, thus limiting the time an access token can be used by an application. The OAuth specification allows the OAuth authorization service to not only grant access tokens but also refresh tokens. Refresh tokens typically have a longer lifetime than access tokens, and applications can use a refresh token to obtain a new access token from the OAuth authorization service as long as the refresh token is still valid. In contrast to obtaining access tokens using the full approval procedure, exchanging a refresh token for a new access token does not require any end-user interaction and especially no end-user authentication. The concept of a refresh token allows authorization of applications to access services on behalf of end users for longer periods of time (refresh token lifetime) while still limiting the exposure to the validity period of an access token by allowing refresh token revocation. If an application at the end of the validity of the currently used access token tries to obtain a new access token, then this requires interaction with the OAuth authorization service; and if the refresh token has been revoked, then the OAuth authorization service simply refuses to issue a new access token representing continued authorization of the application.

OAuth is commonly used with various services in the Internet. Instead of building their own authorization logic into their web applications, some services delegate that to the OAuth authorization services of sites such as Facebook, Google, Twitter, or others. On the main web site the user then only has to click on the icon representing the OAuth authorization link with, for example, Facebook. The main web site (client) will then initiate the OAuth authorization flow, which in turn redirects the end user's user agent (web browser) to the authorization server (for example, Facebook). The end user then authenticates against the authorization server using their credentials, and then the authorization prompts the end user for authorization of the level of access (scopes) the client requested via the flow (for example, access to the user's email address). As soon as the end user grants access, the authorization server grants access and an access token is issued to the client requesting access. The actual process for how the client obtains the access token depends on the type of OAuth authorization flow used by the client.

OAuth Roles

The following role definitions help to explain the operation of OAuth:

- Resource owner or end user

  The owner of the protected resource. In the OAuth framework the resource owner is the entity granting access to the protected resource. This can also be referred to as the end user when the resource owner is a person.

- Resource server

  The protected resource is hosted on this server. Requests to access resources use access tokens as proof of authorization, and the resource server grants or denies access based on the access token provided in the request. Protected resources in the context of the Cisco Collaboration solution are the interfaces used by Cisco Jabber clients and endpoints, including UDS, Unity Connection VMRest, and others.

- Client

  The application making requests to protected resources on behalf of the resource owner. The client can be anything such as an application running on a desktop machine or a mobile device, a server application, or a cloud service. The term "client" only denotes the role in the context of the OAuth framework.

  Depending on the particular use case, an OAuth client may be a Cisco Collaboration service (for example, the Collaboration Edge) or an end user client (for example, Jabber). If the Collaboration Edge requests an OAuth token on behalf of a user, then the Collaboration Edge acts as an OAuth client. In the case of a Jabber client login flow inside the enterprise, the Jabber client acts as the OAuth client.

  Every OAuth client has an unique identifier, the OAuth client_id . This OAuth client_id uniquely identifies a client type. For example, Jabber for Windows and Jabber for Android use different client_ids, but all releases of Jabber for Windows use the same client_id unless concrete reasons mandate a changed client_id to enable the authorization service to differentiate between different client releases (for example, support variation in the OAuth exchange with different client releases). A set of client_ids is predefined for Cisco products and also for a third-party client.

  When requesting authorization to a protected resource, an OAuth client might request a token with a particular scope. The scope indicates the range of services that an OAuth token can be used to access.

  An OAuth Access Token is granted by the authorization service and is used by bearers (clients) for access to a protected resource. Typically access tokens are issued to a specific user and have a specific expiration time. Whenever an access token expires, the client must get a new access token.

- Authorization server

  After authentication of the resource owner and authorization by the resource owner, the authorization server issues access tokens to be used by the client.

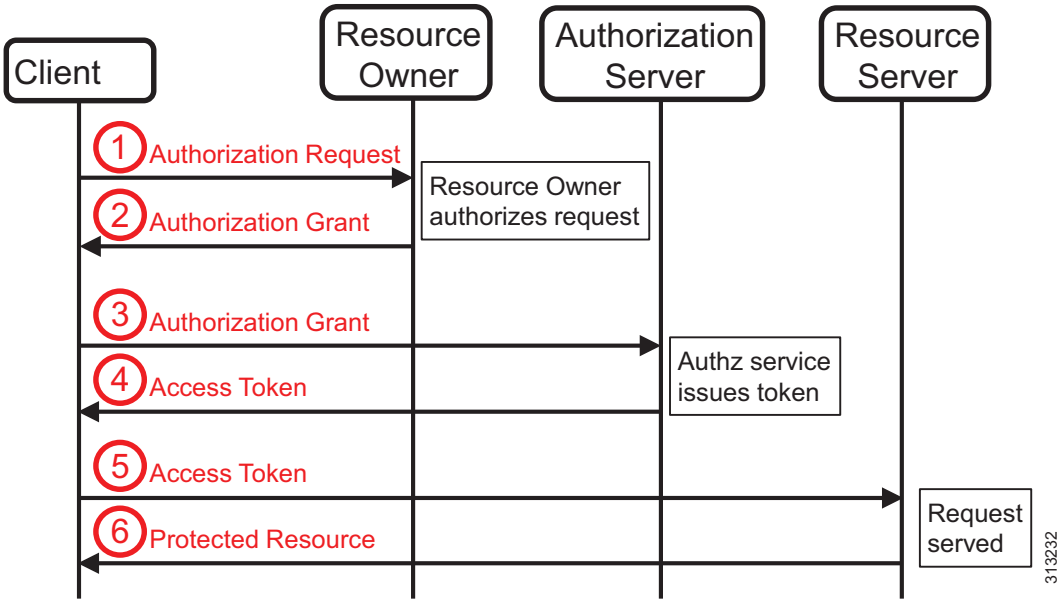
The resource server and the authorization server are not necessarily separate entities; these functions can exist on the same server. In addition, a single authorization server can issue access tokens for multiple resource servers.

In the Cisco Unified Communications architecture the authorization server is a function running on Unified CM, and the access tokens issued by the authorization server are used by Cisco Jabber clients and other endpoints to obtain access to a number of collaboration interfaces (for example, Unified CM SIP, Unified CM CTI, Unified CM IM and Presence SOAP, and Unity Connection VMRest).

General OAuth Flow

OAuth defines a number of different flows to obtain authorization, but all of them share some commonalities which are shown in Figure 16-20.

Figure 16-20 General OAuth Protocol Flow



The OAuth flow shown in Figure 16-20 illustrates the four roles and the following interactions between them:

1. The client requests authorization to access a protected resource from the resource owner (end user). Although Figure 16-20 shows this as a direct interaction between the client and the resource owner, in reality this request typically is made via the authorization server as intermediary. In this case the authorization server, after successful authentication of the resource owner, asks the resource owner to grant authorization to the client that initiated the authorization flow. The representation in Figure 16-20 is to clarify that the authorization of the client lies in the hands of the resource owner (end user).
2. The client receives the authorization grant, a representation of the resource owner's authorization. The grant can be expressed by one of four different grant types defined in the OAuth specification. The grant type depends on the method used by the client to request authorization and the types supported by the authorization server.
3. Using the authorization grant, the client can now request an access token from the authorization server. For this the client needs to authenticate and present the previously acquired authorization grant. Client authentication is required to avoid abuse of the authorization grant through untrusted intermediates.
4. The authorization server authenticates the client, validates the authorization grant, and if valid, issues an access token and (optional) a refresh token. Client authentication by the authorization server requires the client's authentication credentials to be registered with the authorization server beforehand.
5. The client can now request access to a protected resource and use the previously obtained authorization token as proof of authorization.

6. The resource server validates the access token and, if valid, serves the request. This validation can require a transaction between the resource server and the authorization server, especially if self-contained access tokens are not used.

As mentioned in the description of step 1, the preferred method to obtain an authorization grant from the authorization server is by using the authorization server as intermediary. The OAuth authorization code grant flow is an example of this procedure (see the section on Authorization Code Grant Flow, page 16-50, for more details).

Authorization Grants

As shown by the description of the general OAuth authorization flow in Figure 16-20, an authorization grant is a credential representing the authorization to access a protected resource granted by the resource owner. The authorization grant is used by the client to obtain an access token. The OAuth specification defines four grant types: authorization code, implicit, resource owner password credentials, and client credentials. For the purpose of this document only two flows are relevant:

Implicit Grant Flow

The implicit grant is a simplified authorization code grant flow. Instead of issuing an authorization code grant, the authorization server directly issues an access token to the client. This is called "implicit" because no intermediary credentials are issued.

This flow is optimized for clients implemented in a browser using scripting languages. The client is not authenticated in this flow, since the access token is issued directly. This exposes the access token to the resource owner and potentially other applications having access to the resource owner's browser.

While the implicit grant flow is more responsive (no additional transaction to exchange the authorization code grant for an access token), the implicit grant flow should be considered less secure.

Also in the context of Cisco Unified Communications solutions, where access tokens obtained via OAuth authorization flows are used by Jabber clients and other endpoints to access various system interfaces, it is important to note that to obtain a new access token when the reaching the lifetime of the previous access token with the implicit grant flow always a new authentication step is required while with the authorization code grant flow the client also obtains a refresh token which can be used to obtain a new access token directly without additional end user authentication.

SAML Bearer Assertion Grant Flow

An entity (typically a service) uses an assertion issued on behalf of an end user to get an OAuth token that is associated with the end user. A variation of this flow is used by the Collaboration Edge to get tokens on behalf of clients connecting from outside the edge to obtain authorization codes.

Authorization Code Grant Flow

This flow uses the authorization server as intermediary between the client and the resource owner, as shown in Figure 16-21.

Figure 16-21 Authorization Code Grant Flow

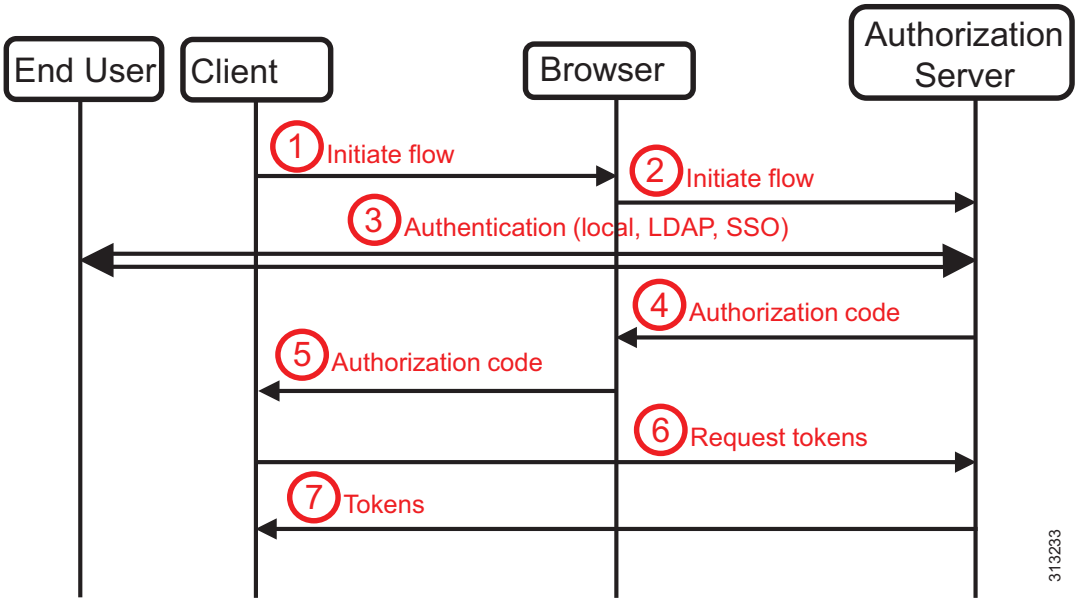


Figure 16-21 shows the details of the authorization code grant flow. The client does not directly request the authorization grant (authorization code) from the end user. Instead the client directs the end user to an authorization server which in turn later in the flow directs the end user back to the client with the authorization code. For this redirection the end user's web browser is used, and the authorization code grant flow is browser based. For the redirection of the end user's browser in this flow, HTTP 302 redirects can be used as well as JavaScript code-based redirection in web content returned to the browser.

1. The client (for example Cisco Jabber) initiates the authorization flow by redirecting the end user's browser to the authorization endpoint. For the Cisco Unified Communications solution this is /ssosp/oauth/authorize on Unified CM.
2. The browser accesses (GET) the endpoint on the authorization server. With this request a number of parameters are passed: the client identifier, the requested level of access (scopes), a unique request identifier, and the redirection URI to which the resulting authorization code should be posted at the end of the flow.
3. The authorization server now authenticates the end user. For authentication against the local end user table on Unified CM or using LDAP bind, this involves asking the end user to enter their username and password. For this authentication based on username and password, the authorization server returns a web form as the result to the GET on the authorization endpoint, the end user enters the credentials, and the authorization server validates the credentials either against the local end user table or against the configured LDAP server using an LDAP bind.

   If SSO is configured, the authorization server initiates a SAML 2.0 Redirect/Post flow. Posting the SAML assertion to the authorization service at the end of the Redirect/Post flow finishes the authentication step in this case.

After successful authentication, the Cisco authorization service immediately proceeds to issues the authorization code grant; the end user's authorization of the client (for example, Cisco Jabber) to use the requested resources is assumed as given.

4. To grant the authorization code to the client, the authorization server redirects the end user's browser to the redirect URI provided in steps 1 and 2. The redirection URI includes the authorization code and the request identified from step 2.
5. The browser accesses the URL and thus reveals the authorization code and request identified to the client. The request identified allows the client to correlate this event with the outstanding authorization event that triggered the flow.
6. The client now requests an access token from the authorization service using the authorization code obtained in the previous step. The token request to the authorization service is authenticated using the client's credentials (client ID and secret), and the client also again passes the redirection URI.
7. The authorization server authenticates the client, checks the redirection URL, and if valid responds back with an access token and a refresh token.

It is important to note that, because end-user authentication is obtained by the authorization server using local authentication, LDAP bind, or SSO, resource owner authentication details (for example, type of authentication and resource owner credentials) are never shared with the client. The client obtains only the authorization code grant.

Other benefits of this flow include:

- The access token is obtained by the client directly and thus is not exposed to the resource owner’s browser.
- The transaction to exchange the authorization code grant for an access token is authenticated using client credentials. The authorization code grant alone cannot be used to obtain access to the protected resource without knowledge of the client credentials.

The authorization code grant flow was introduced with Cisco Unified CM release 11.5(1) SU3. The authorization code grant flow with refresh tokens needs to be via the enterprise parameter **OAuth with Refresh Login Flow**. The default setting for this parameter is "disabled," but Cisco recommends enabling this flow. The implicit grant flow is always supported for backward compatibility even if the authorization code grant flow with refresh tokens is enabled.

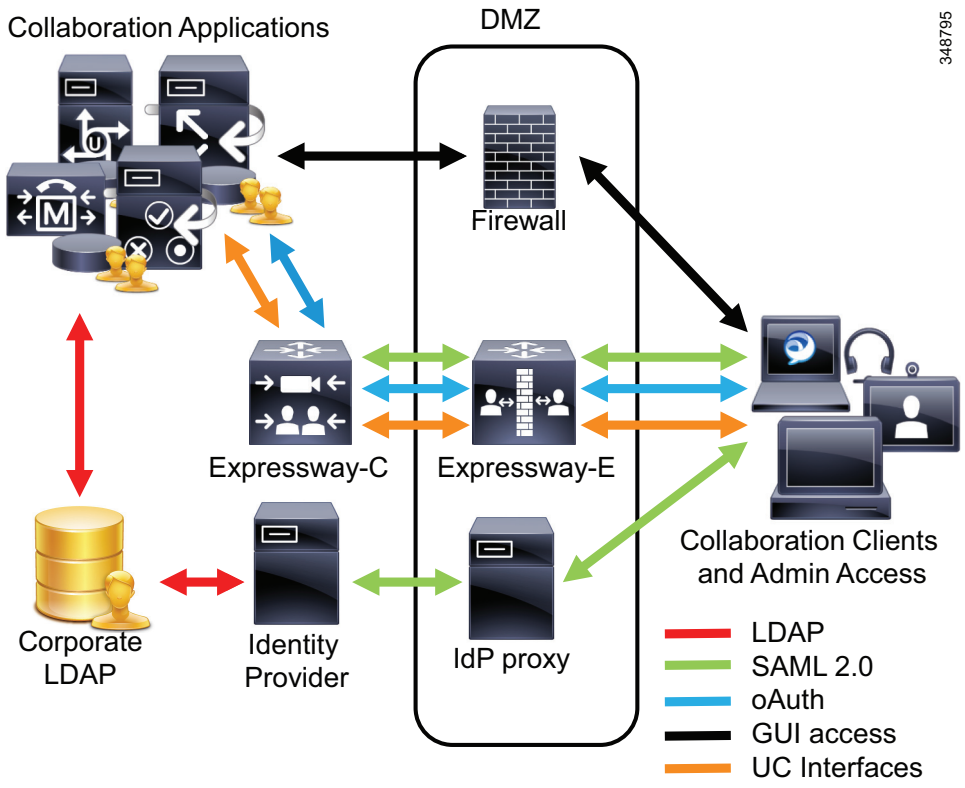
Tokens

Access tokens used in the Cisco Collaboration solution have a default lifetime of one hour (3,600 seconds). If an expired access token is used, then the resource server rejects the service and returns an error indicating that the token has expired. The client then has to obtain a new access token. To avoid this forced access token refresh, Cisco Collaboration clients initiate a token refresh after 75% of the token expiration time has elapsed. Note that access token expiration does not affect existing sessions with protocols such as XMPP and SIP, where authorization based on access tokens happens only during session establishment (for example, during SIP registration).

Mobile and Remote Access (MRA) Authentication and Authorization

To provide OAuth authorized access for clients outside the enterprise that are connected using mobile and remote access (MRA) through the Collaboration Edge, deploy Cisco Expressway-E and Expressway-C as illustrated in Figure 16-22.

Figure 16-22 OAuth with Collaboration Edge



Endpoints and collaboration clients registering through the Collaboration Edge can use OAuth to obtain access tokens by using the authorization API on Expressway-C (proxied by Expressway-E). When calling this authorization endpoint, the client needs to specify that the authorization code grant flow should be used along with the client_id, a request-specific state identifier, and the user identification. The user can be identified by username, email address, or user identifier. The authorization API on Expressway-C then will either initiate a SAML Redirect/Post authentication flow by redirecting the browser to an IdP or present a local authentication page, depending on the authentication method of the user. The authentication method for the user is determined by the MRA access control settings on Expressway-C and the home cluster settings of the user:

- The **Authentication path** setting reflects whether authentication via Unified CM LDAP, SAML SSO, or both should be allowed.
- **Authorize by OAuth token with refresh** enables the authorization code grant flow with refresh tokens.
- **Authorize by OAuth token** enables the legacy implicit grant authorization flow. This is not needed with Unified CM 12.*x* and current Jabber clients.

- **Authorize by user credential** needs to be enabled to allow MRA functionality for all IP phones and Cisco TelePresence endpoints.
- **Check for internal authentication availability** needs to be enabled only until all Unified CM clusters are migrated to a release of Unified CM that supports OAuth with refresh token and all clusters are using a common authentication (either SSO or Unified CM LDAP basic authentication). When this parameter is enabled, Expressway-C will first determine a user's home Unified CM cluster and then determine the authentication settings on the home cluster.

Depending on the required authentication method, the respective authorization flow is initiated.

MRA Sign-On with Local Authentication

Figure 16-23 shows the flow that is used to obtain an access token with authentication through Cisco Expressway based on username and password.

Figure 16-23 OAuth Authorization through Expressway with Local Authentication

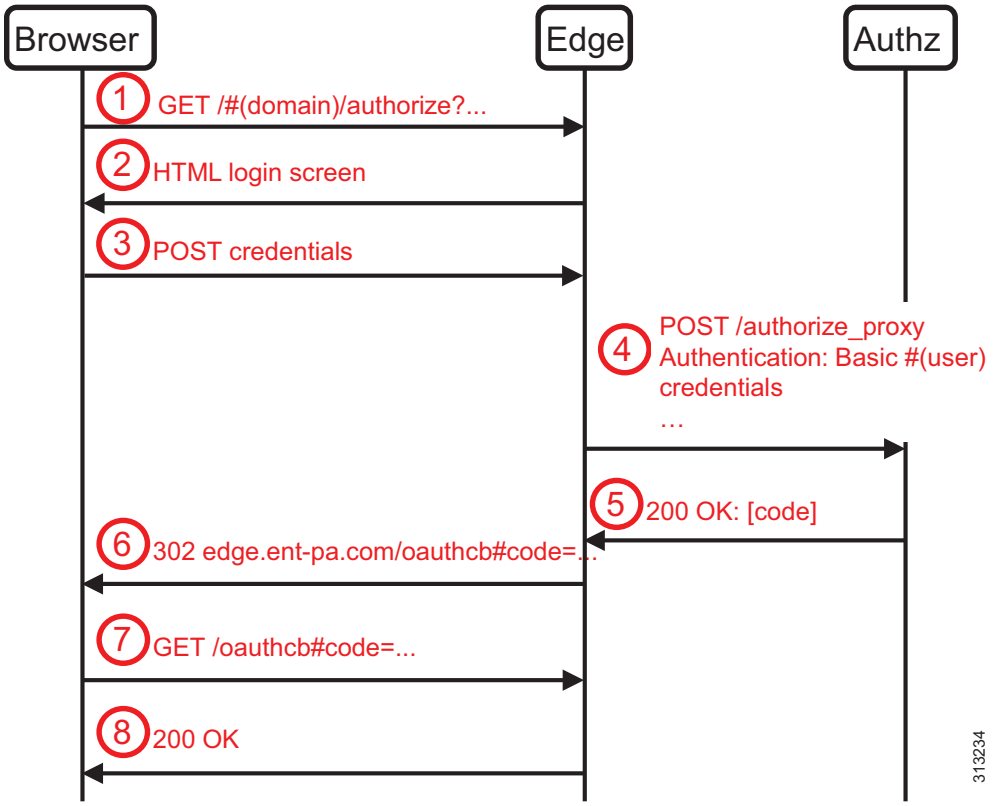


Figure 16-23 illustrates the following flow steps:

1. The Jabber client directs the web browser to access the authorization API on Expressway. The request contains state information uniquely identifying the request, the client_id, and the user identification.
2. Based on the user identification, Expressway determines that username and password authentication is required and then returns a web page with a web form to enter the user credentials. All parameters obtained in step 1 are passed through hidden input fields in the web form.

3. After the user credentials are entered, the web form is posted back to Expressway.
4. Expressway uses the /authorize_proxy endpoint on the authorization service on Unified CM to obtain the authorization code. This is a variation of a SAML bearer assertion grant flow. This request is authenticated using the username and password of an application user. The referenced application user has to have rights to access the AXL API on the authorization service on Unified CM. The /authorize_proxy request contains all authorization parameters cached earlier.
5. The authorization service validates the credentials by checking them against the local end-user table or via an LDAP bind. The authorization service then returns the authorization code to (the Collaboration) Edge.
6. Edge can then cache the code but also needs to return the code to the client. This is achieved by returning a 302 message to the client, redirecting the browser instance on the client to the OAuth callback on Edge. The target URL of the redirect contains the required information about the authorization code.
7. The browser on the client follows the redirection and accesses the OAuth callback resource on Edge.
8. The 200 OK message finishes the SSO flow through Edge. The client can now extract the authorization code from the final URL.

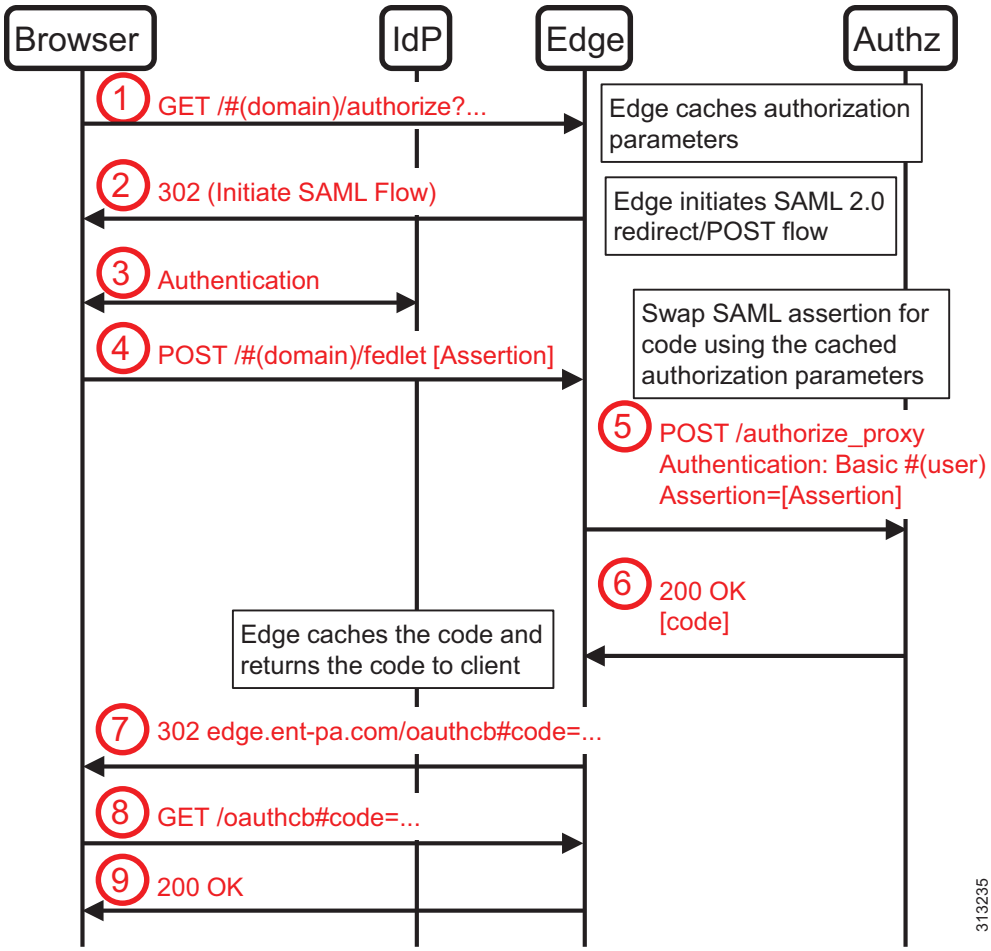
As a result, the authorization parameters are now cached on Expressway, and the Jabber client owns an authorization code.

Jabber can exchange the authorization code for an access token by again accessing the authorization API on Expressway. Expressway in this case, similar to steps 4 and 5 in the above flow, proxies the request to the access_token endpoint of the authorization service.

MRA Sign-On with SSO Authentication

This flow is very similar to the flow for authentication based on username and password. In this case Cisco Expressway uses a SAML 2.0 Redirect/Post flow to obtain authentication, and the browser on the client is redirected to a publicly accessible identity provider. This typically is an IdP proxy in the customer's DMZ, acting as proxy for the IdP deployed inside the enterprise. The IdP proxy in the DMZ essentially is only a generic HTTPS reverse proxy for the enterprise IdP. Some IdP vendors offer an option to install an IdP instance in the DMZ as an IdP proxy role. Expressway-E and Expressway-C proxy only proxy collaboration client requests for services on the collaboration applications. While the SAML authentication flow is redirected to an IdP proxy in the DMZ by making sure that the public DNS resolved the DNS name of the enterprise IdP to the public IP address of the IdP proxy in the DMZ, the OAuth exchange to achieve an OAuth token passes through Expressway-C and Expressway-E, and Expressway-E requests the OAuth token as a proxy for the actual client. This is a variant of an OAuth SAML bearer grant flow as shown in Figure 16-24.

Figure 16-24 OAuth Authorization through Expressway with SSO Authentication



1. To acquire an OAuth token, the browser sends an HTTP GET request to the /authorize endpoint on Edge. The /authorize endpoint on Edge is accessed using prefix encoding to refer to the customer domain. Edge in this description refers to a Cisco Expressway-C and Expressway-E pair implementing the Collaboration Edge.
2. Expressway initiates an SP-initiated SAML 2.0 redirect/POST authentication flow by returning a 302 response redirecting the browser to the IdP. Edge also caches the authorization parameters from the client request because they are needed later in the actual OAuth proxy request.
3. Browser and IdP then exchange the messages required to authenticate the user. The message exchange depends on the authentication method configured on the IdP.
4. If the SAML authentication succeeds, then as the last step of the SAML exchange the browser POSTs the SAML assertion to the assertion consumer service on Edge. Edge still needs to exchange this SAML assertion for an authorization code.
5. To achieve this, Edge uses the /authorize_proxy endpoint on the authorization service. This request is authenticated using the username and password of an application user. The referenced application user has to have rights to access the AXL API on the authorization service on Unified CM. The /authorize_proxy request contains all authorization parameters cached earlier.

6. The authorization service then can check whether the authenticated end user has the required privileges. If the authenticated user is authorized to access the requested service, then the authorization service issues an authorization code and returns that code in a 200 OK message.
7. Edge can then cache the code and still needs to return the code to the client. This is achieved by returning a 302 message to the client, redirecting the browser instance on the client to the OAuth callback on Edge. The target URL of the redirect contains the required information about the authorization code.
8. The browser on the client follows the redirection and accesses the OAuth callback resource on Edge.
9. The 200 OK message finishes the SSO flow through Edge. The client can then extract the authorization code from the final URL.

As a result, the authorization parameters are cached on Expressway, and the Jabber client owns an authorization code.

Jabber can exchange the authorization code for an access token by again accessing the authorization API on Expressway. Expressway in this case, similar to the flow above, proxies the request to the access_token endpoint of the authorization service.

Understanding OAuth Tokens

This sections covers access and refresh tokens, token expiration, and token management.

Access Tokens

Access tokens are issued to clients by the authentication service. The content of the access token is opaque to the client. Clients do not need to understand the semantics of access tokens; they are treated as an arbitrary string value by the client. Access tokens are used only to authorize requests sent to services by the clients.

Access tokens issued by the authorization service on Cisco Unified CM are self-contained tokens. These self-contained tokens can be validated by Unified Communications services without contacting the authorization service. The access tokens are JSON Web Tokens as described in RFC 7519, and they contain information about the authorized user, the expiration, and the authorized scopes. The access tokens are encrypted and digitally signed by the authorization service.

Because Unified Communications services check authorization based on self-contained access tokens without contacting the authorization service, there is no way to centrally revoke previously authorized access within the validity period of an access-token. To limit the exposure, self-contained access tokens typically have a short lifetime. The default lifetime of access tokens issued by the authorization service on Cisco Unified CM is 60 minutes, and it can be set to values between 1 minute and 1440 minutes (one day).

Jabber typically tries to obtain a new access token as soon as the current access token has 25% lifetime left; thus, a token with a lifetime of 60 minutes will be refreshed after 45 minutes.

Refresh Tokens

Refresh tokens are issued to clients by the authorization service on Cisco Unified CM. Clients can present a refresh token to the authorization service on Unified CM to obtain an access token within the lifetime of the refresh token. This does not require end-user authentication, so that obtaining a new access token is a very fast transaction that can be executed without impacting the user experience. Again, the content of a refresh token is opaque to the client. The request to obtain a new access token is authenticated using the client credentials (client ID, client secret, and redirect URI).

The default lifetime of refresh tokens issued by the authorization service on Unified CM is 60 days, and can be set to values between 1 day and 1825 days (five years).

Note Changing the refresh token lifetime invalidates all refresh tokens currently issued by the authorization service, thereby requiring all users to re-authenticate to obtain new refresh tokens.

A new full authorization flow is required to obtain a new refresh token. This requires end-user authentication. Jabber clients will advise end users when refresh tokens are about to expire, and a new authorization flow can then be initiated by the end user.

Administrators can revoke all refresh tokens for a user by means of the **https://<*unified_cm*>:8443/ssosp/token/revoke?user_id=<*uid*>** endpoint on Unified CM. Here *unified_cm* needs to be replaced with the IP address of hostname of the Unified CM publisher and *uid* needs to be replaced with the user id of the user whose refresh tokens are to be revoked. The request needs to be authenticated using the credentials of an administrator user.

Token Signing and Encryption Keys

The self-contained access tokens are encrypted and digitally signed by the authorization service on Cisco Unified CM. The required keys are created on the Unified CM publisher node and are distributed across all the nodes of the cluster. While Unified CM IM and Presence obtains the keys via intra-cluster replication, Cisco Expressway and Unity Connection need to pull the keys from Unified CM to enable access token validation. Access to the keys is obtained via a token key API on Unified CM. Access to this API requires authentication using credentials of an application user with AXL access. On Cisco Unity Connection, authorization servers are defined in the **Authz Server** section in the **System Setting** tab in Cisco Unity Connection Administration.

Signing and encryption keys can be regenerated if the administrator believes that the keys have been compromised. Regenerating either of these keys invalidates all access tokens issued by the authorization service, so that all clients need to obtain new tokens leading to re-authentication of all end users.

Signing keys can be regenerated using the **set key regen authz signing** CLI command. Encryption keys can be regenerated using the **set key regen authz encryption** CLI command. Information about the current signing and encryption keys can be displayed using the **show key authz signing** and **show key authz encryption** CLI commands.

Scopes

Access tokens contain a scope element. The scope defines the Unified Communications services that the holder of the access token is authorized to use. The scopes for access tokens issued for any given user are defined by setting the **Jabber Desktop Client Policy** and **Jabber Mobile Client Policy** under **Mobile and Remote Access Policy** in the user profile configuration on Cisco Unified CM. This allows the administrator to define different scopes for Jabber desktop and Jabber mobile clients. Possible values are **No Service**, **IM&Presence only**, and **IM&Presence, Voice and Video calls**.

The scope is checked by Expressway whenever a client establishes a connection and Expressway will only establish connections to authorized services.

PART 3

Collaboration Applications and Services

Contents of This Part

This part of the document contains the following chapters:

- Overview of Collaboration Applications and Services
- Cisco Unified CM Applications
- Cisco Voice Messaging
- Collaboration Instant Messaging and Presence
- Mobile Collaboration
- Cisco Unified Contact Center
- Call Recording and Monitoring

Overview of Collaboration Applications and Services

Revised: June 15, 2015

Once the network, call routing, and call control infrastructure has been put in place for your Cisco Unified Communications and Collaboration System, additional applications and services can be added or layered on top of that infrastructure. There are numerous applications and services that can be deployed on an existing Cisco Unified Communications and Collaboration infrastructure, and the following applications and services are typically deployed:

- Cisco Unified Communications Manager applications — Provide enhanced features and functionality for IP telephony.
- Voice messaging — Provides voicemail services and message waiting indication.
- Presence services — Provide user availability tracking across user devices and clients.
- Mobility services — Provide enterprise-level Unified Communications and Collaboration features and functionality to users outside the enterprise.
- Contact center — Provides call handling, queuing, and monitoring for large call volumes.
- Call recording — Provides the ability to record audio and video calls for later retrieval and playback.

The chapters in this part of the SRND cover the applications and services mentioned above. Each chapter provides an introduction to the application or service, followed by discussions surrounding architecture, high availability, capacity planning, and design considerations. The chapters focus on design-related aspects of the applications and services rather than product-specific support and configuration information, which is covered in the related product documentation.

This part of the SRND includes the following chapters:

- Cisco Unified CM Applications, page 18-1

  This chapter covers Cisco Unified Communications Manager (Unified CM) applications, which provide numerous operational and functional enhancements to basic IP telephony. External eXtensible Markup Language (XML) productivity applications or IP Phone Services can run on the web server and/or client on most Cisco Unified IP Phones. This chapter also discusses a number of Unified CM integrated applications that provide additional functionality, such as Cisco Extension Mobility, Cisco Unified Communications Manager Assistant, and Cisco WebDialer.

- Cisco Voice Messaging, page 19-1

  This chapter examines voice messaging, a common and prevalent application within most Unified Communications and Collaboration deployments, which allows callers to send messages and allows subscribers of the system to retrieve messages. The chapter examines messaging deployment models, voice messaging features and functionality, voicemail networking, and design and deployment best practices for voice messaging applications.

- Collaboration Instant Messaging and Presence, page 20-1

  This chapter discusses presence services, an increasingly critical piece of most Unified Communications and Collaboration deployments due to the productivity improvements that can be realized from user availability-based applications. This chapter defines presence and explores the various presence components and features, protocols, deployment models, redundancy, capacity, and general design guidelines.

- Mobile Collaboration, page 21-1

  This chapter looks at mobility applications, which are becoming extremely important given the growth of mobile work forces and the blurring of enterprise boundaries for Unified Communications and Collaboration features and services, resulting in an increased demand for mobility applications and services. This chapter discusses mobility solution architectures, functionality, and design and deployment considerations.

- Cisco Unified Contact Center, page 22-1

  This chapter covers contact center solutions, an important and integral part of large Unified Communications and Collaboration deployments that require high-volume call center applications. This chapter examines call center solution architectures, functionality, and design and deployment implications.

- Call Recording and Monitoring, page 23-1

  This chapter provides an overview of various call recording and monitoring solutions available for Cisco Unified Communications and Collaboration systems for both audio and video calls. The chapter also outlines basic design considerations for call recording and monitoring solutions embedded within a Cisco Unified Communications and Collaboration solution.

Architecture

As with other network and application technology systems, Unified Communications and Collaboration applications and services must be layered on top of the underlying network and system infrastructures. Unified Communications and Collaboration applications and services such as voice messaging, rich media conferencing, presence, mobility, contact center, and call recording rely on the underlying Unified Communications and Collaboration call routing and call control infrastructure and network infrastructure for everything from network connectivity to basic Unified Communications and Collaboration functions such as call control, supplementary services, dial plan, bandwidth management, and gateway services. For example, voice messaging and presence applications leverage the network infrastructure to reach users in campus sites, in branch sites, and on the Internet. Further, these same applications depend on the Unified Communications and Collaboration voice and video endpoints, call routing, PSTN connectivity, and media resources provided by the call routing and control infrastructure. In addition to relying on these infrastructure layers and basic Unified Communications and Collaboration services, applications and services are also often dependent upon each other for full functionality.

High Availability

As with network, call routing, and call control infrastructures, critical Unified Communications and Collaboration applications and services should be made highly available to ensure that required features and functionality remain available if failures occur in the network or applications. It is important to understand the various types of failures that can occur and the design considerations around those failures. In some cases, the failure of a single server or feature can impact multiple services because many Unified Communications and Collaboration applications are dependent on other applications or services. For example, while the various application service components of a contact center deployment might be functioning properly, the loss of all call control servers would effectively render the contact center unusable because the deployment is dependent upon the call control servers to route calls to the call center applications.

For applications and services such as voice messaging and mobile collaboration, high availability considerations include temporary loss of functionality due to network connectivity or application server failures resulting in the inability of callers to leave messages, of users to retrieve messages, and of users to schedule or attend conferences. In addition, failover considerations for callers and users of voice messaging and mobile collaboration applications include scenarios in which portions of the functionality can be handled by a redundant resource that allows end users to continue to access services in the event of certain failures.

High availability considerations are also a concern for services such as presence and mobility. Interrupted network connectivity or server failures will typically result in reduced functionality or, in some case, complete loss of functionality. For presence services, this can mean that some or all devices and clients will be unable to send or receive presence or availability updates. For mobility services, high availability considerations include the potential for loss of specific functionality such as two-stage dialing or dial-via-office, or reduced functionality for features such as single number reach (resulting in situations where only the enterprise phone rings or only the mobile phone rings). Further, in some failure scenarios, enterprise endpoints and mobile clients might have to re-register, re-connect, and/or re-authenticate before full functionality is available again.

For contact center deployments, there are numerous servers and components for which high availability must be considered. Typically, an isolated single-server or single-component failure can be handled without loss of features or functionality as long as the server or component has been made redundant. In other situations, loss of multiple servers or components will typically result in loss of some features or functionality. In scenarios where there is complete loss of a particular component such as all call control servers, more catastrophic loss of features or functionality is possible.

When considering collaboration clients and applications, high availability is certainly important. Not only can specific collaboration features or functions become unavailable in failure scenarios, but in some cases presence-capable clients might be unable to connect to the network for even basic functionality such as registration and making or receiving calls. In other cases, clients or devices might have to reconnect and re-authenticate in order to return to service.

Capacity Planning

Network, call routing, and call control infrastructures must be designed and deployed with an understanding of the capacity and scalability of the individual components and the overall system. Similarly, deployments of Unified Communications and Collaboration applications and services must also be designed with attention to capacity and scalability considerations. When deploying various Unified Communications and Collaboration applications, not only is it important to consider the scalability of the applications themselves, but you must also consider the scalability of the underlying infrastructures. Certainly the network infrastructure must have available bandwidth and be capable of handling the additional traffic load the applications will create. Likewise, the call routing and control infrastructure must be capable of handling user and device configuration and registration as well as application integration loads surrounding protocols and connections. For example, with applications and services such as mobility, presence, and contact center, there are capacity implications for each of these individual applications in terms of users, devices, and features, but just as important is the scalability of the underlying infrastructure to handle connections and protocols such as Computer Telephony Integration (CTI). While a mobility, presence, or contact center application may be able to support many CTI connections, the underlying call control and routing infrastructure might not have available capacity to handle the added CTI load of the application or service.

For applications and services such as voice messaging and rich media conferencing, capacity planning considerations include things like number of mailboxes or users, mailbox size, audio and video ports, and MCU sessions. In most cases additional capacity can be added by increasing the number of application servers and MCUs or by upgrading server or MCU hardware with higher-scale models, assuming the underlying network and call routing and control infrastructures are capable of handling the additional load.

Capacity planning considerations are also a concern for services such as presence and mobility. Scalability must be contemplated not only for things like numbers of configured and supported users and devices, but also for the number of integrations and connections between the applications and services. The volume of two-stage dialing and dial-via-office calls is of particular concern for mobility applications from the perspective of both the call control capacity and the PSTN gateway capacity. With presence services, on the other hand, critical scalability concerns include frequency of presence status changes and the propagation of those changes to the network, as well as text or instant message volumes. Typically, additional application servers or hardware upgrades will result in increased capacity for the applications and services, but the underlying call routing and control infrastructures must be capable of handling any increases in load.

Contact center deployments are no different than other applications and services in terms of scalability concerns. Certainly the number of agents and agent devices handling calls is important in terms of user and device configuration and registration. However, the major concerns in terms capacity for contact center deployments are the high number of busy hour call attempts (BHCA) common in contact centers and the number of CTI integrations to the call control and routing infrastructure.

When considering collaboration clients and application capacity planning, device registration and configuration are the most important scalability concerns. However, there are other scalability implications in terms of the back-end applications and services such as presence and messaging. Further, when deploying or integrating various clients with third-party applications and infrastructures, you must also consider the supported capacities for those third-party deployments.

For a complete discussion of system sizing, capacity planning, and deployment considerations related to sizing, refer to the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Cisco Unified CM Applications

Revised: March 1, 2018

Cisco Unified Communications Manager (Unified CM) applications provide numerous operational and functional enhancements to basic IP telephony. External eXtensible Markup Language (XML) productivity applications or IP Phone Services can be run on the web server and/or client on most Cisco Unified IP Phones. For example, the IP phone on a user's desk can be used to get stock quotes, weather information, flight information, and other types of web-based information. In addition, custom IP phone service applications can be written that allow users to track inventory, bill customers for time, or control conference room environments (lights, video screen, temperature, and so forth). Unified CM also has a number of integrated applications that provide additional functionality, including:

- Cisco Extension Mobility (EM)

  The Extension Mobility feature enables mobile users to configure a Cisco Unified IP Phone as their own, on a temporary basis, by logging in to that phone.

- Cisco Unified Communications Manager Assistant (Unified CM Assistant)

  Unified CM Assistant is a Unified CM integrated application that enables assistants to handle one or more managers' incoming phone calls.

- Cisco WebDialer

  WebDialer is a click-to-call application for Unified CM that enables users to place calls easily from their PCs using any supported phone device.

In some cases these integrated applications also invoke IP Phone Services to provide additional functionality.

This chapter examines the following Unified CM applications:

- IP Phone Services, page 18-2
- Extension Mobility, page 18-7
- Unified CM Assistant, page 18-19
- WebDialer, page 18-34

This chapter also covers:

- Cisco Unified Attendant Consoles, page 18-42
- Cisco Paging Server, page 18-47

What's New in This Chapter

Table 18-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 18-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Cisco Unified Attendant Consoles	Cisco Unified Attendant Consoles, page 18-42	March 1, 2018
Other minor updates and corrections	Various sections of this chapter	March 1, 2018

IP Phone Services

Cisco Unified IP Phone Services are applications that utilize the web client and/or server and XML capabilities of the Cisco Unified IP Phone. The Cisco Unified IP Phone firmware contains a micro-browser that enables limited web browsing capability. These phone service applications provide the potential for value-added services and productivity enhancement by running directly on the user's desktop phone. For purposes of this chapter, the term *phone service* refers to an application that transmits and receives content to and from the Cisco Unified IP Phone.

This section examines the following design aspects of the IP Phone Services feature:

- IP Phone Services Architecture, page 18-2
- High Availability for IP Phone Services, page 18-5
- Capacity Planning for IP Phone Services, page 18-6
- Design Considerations for IP Phone Services, page 18-7

IP Phone Services Architecture

An IP Phone service can be initiated in several ways:

- User-initiated (pull)

  An IP Phone user presses the Services or Applications button, which sends an HTTP GET message to Unified CM for displaying a list of user-subscribed phone services. Figure 18-1 illustrates this functionality.

- Phone-initiated (pull)

  An idle time value can be set within the IP Phone firmware, as indicated by the URL Idle Time parameter. When this timeout value is exceeded, the IP Phone firmware itself initiates an HTTP GET to the idle URL location specified by the URL Idle parameter.

- Phone service-initiated (push)

  A phone service application can push content to the IP Phone by sending an HTTP POST message to the phone.

Note Unlike with the user-initiated and phone-initiated pull functionality, whereby the phone's web client is used to invoke phone services, the phone service-initiated push functionality invokes action on the phone by posting content (via an HTTP POST) to the phone's web server (not to its client).

Figure 18-1 shows a detailed illustration of the user-initiated IP Phone service operation. With Services Provisioning set to **External URL** or **Both** when a user presses the Services or Applications button, an HTTP GET message is sent from the IP Phone to the Unified CM getservicesmenu.jsp script by default (step 1). You can specify a different script by changing the Phone URL enterprise parameter. The getservicesmenu.jsp script returns the list of phone service URL locations to which the individual user has subscribed (step 2). The HTTP response returns this list to the IP Phone (step 3). Any further phone service menu options chosen by the user continue the HTTP messaging between the user and the web server containing the selected phone service application (step 4).

By default the Services Provisioning parameter is set to **Internal**. With this setting, the IP phone obtains the list of phone services from its configuration file instead of sending an HTTP GET message to Unified CM.

Note If the Service Provisioning enterprise parameter is set to Internal, steps 1 through 3 are bypassed and the operation of phone services begins with step 4.

Note The Cisco Unified IP Phone 7960 does not have the ability to parse the list of phone services from its configuration file, so it sends an HTTP GET to Unified CM to get that list, even if the Service Provisioning enterprise parameter is set to **Internal**.

Figure 18-1 User-Initiated IP Phone Service Architecture

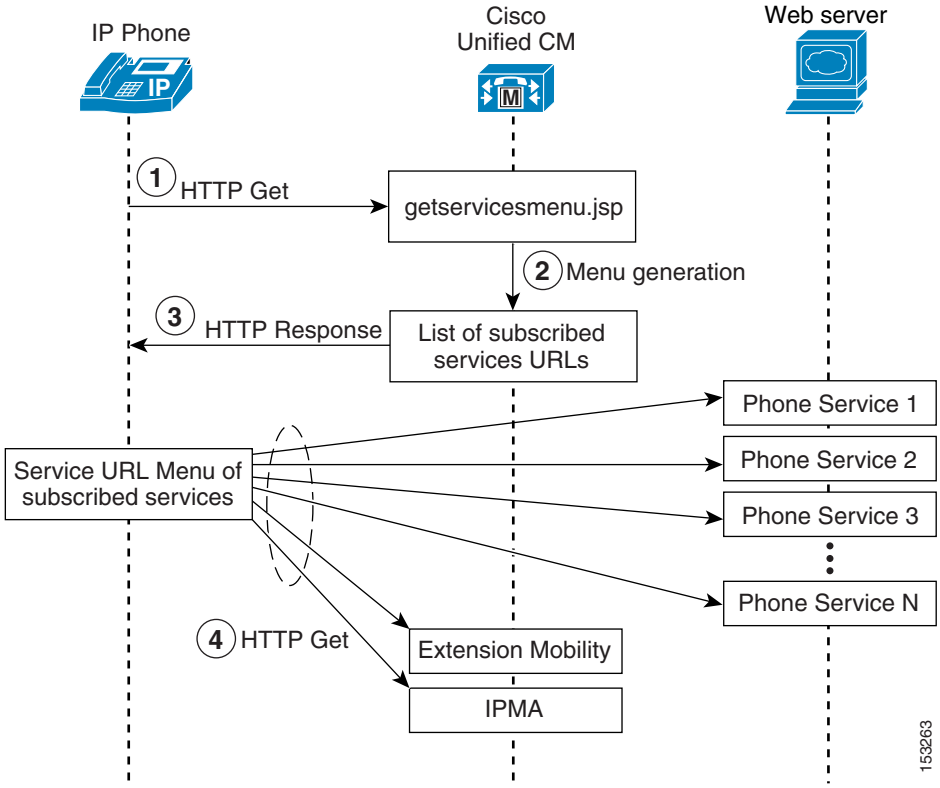
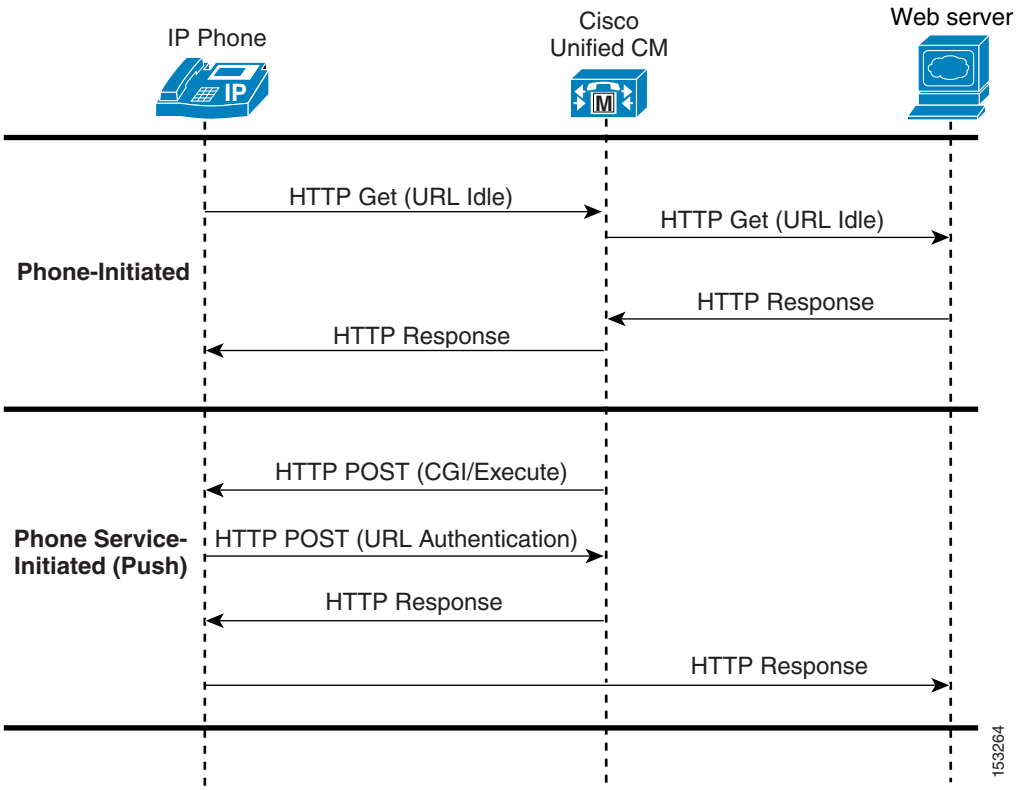


Figure 18-2 shows examples of both phone-initiated and phone service-initiated push functionality. In the phone-initiated example, the phone automatically sends an HTTP GET to the location specified under the URL Idle parameter when the URL Idle Time is reached. The HTTP GET is forwarded via Unified CM to the external web server. The web server sends back an HTTP Response, which is relayed by Unified CM back to the phone, and the phone displays the text and/or image on the screen.

In the phone service-initiated push example, the phone service on the external web server sends an HTTP POST with a Common Gateway Interface (CGI) or Execute call to the phone's web server. Before performing the CGI or Execute call, the phone authenticates the request using the proxy authentication service specified by the URL Authentication parameter. This proxy authentication service provides an interface between the phone and the Unified CM directory in order to validate requests made directly to the phone. If the request is authenticated, Unified CM forwards an HTTP Response to the phone. The phone's web server then performs the requested action, and the phone returns an HTTP response back to the external web server. If authentication fails, Unified CM forwards a negative HTTP Response, and the phone does not perform the requested CGI or Execute action but in turn forwards a negative HTTP Response to the external web server.

Figure 18-2 Phone-Initiated and Phone Service-Initiated IP Phone Service Architecture



In addition to XML Services, a new service can be created with a Service Category of Java MIDlet. When a Java MIDlet-type service is invoked, the configured Service URL contains the URL from which the MIDlet JAD file can be retrieved. When the application server receives the JAD file request, the server should return the appropriate JAR file for that device, which the phone's MIDlet-installer will download and process.

For more information on Java MIDlet support on Cisco IP Phones, refer to the Cisco IP Phone data sheets at https://www.cisco.com.

Note After a phone has downloaded its configuration file, the phone parses the services configuration to determine whether or not the list of services has changed, and if so, it updates its local (persisted) services configuration. If any of the changed services were Java MIDlets (which are explicitly provisioned and stored on the phone), then the phone sequentially walks through the necessary install, upgrade, downgrade, and uninstall operations to comply with what was provisioned in the configuration file. If a MIDlet install fails, it will re-attempt the install the next time the phone checks its configuration file (during boot, reset, or restart).

The administrator has the added ability to specify the Service Type of configured services to be one of the following: IP Phone Services, Directories, or Messages. This gives the administrator the flexibility to control which button users must press on the IP phone to access new services. New services can optionally be configured as Enterprise Subscriptions, which forces them to appear automatically on all IP phones without the need to update subscriptions for each individual phone. In addition, services can be enabled or disabled without the need to delete the service from the Unified CM database.

Note Default services such as Missed Calls, Placed Calls, and Corporate Directory can also be disabled. This allows the administrator to create a custom service with a Service URL matching that of the corresponding default service, thus allowing phones to subscribe to these default services on an as-needed basis.

Unified CM provides the ability to configure a secure IP Phone Services URL using HTTPS in addition to a non-secure URL. Phones that support HTTPS will automatically use the secure URL. For more information about Trust Verification Services and security certificate handling for IP phones, along with a complete list of phones that support HTTPS, refer to the HTTPS information in the latest version of the *Security Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

High Availability for IP Phone Services

To ensure reliable services for phone users, you must maintain a high level of system availability, with a seamless transition to redundant systems during a system failure.

With Services Provisioning set to Internal, the phone will receive its subscribed phone services from the phone's configuration file and store these (and their corresponding service URLs) in flash. This allows the phone to access the service URLs directly on a web server without first querying the Cisco CallManager IP Phone Service. With Services Provisioning set to Internal, the Corporate and Personal Directories default services also have an extra level of redundancy built into the phones. When these services are selected, the phone will attempt to send an HTTP message with the proper URL string to the Unified CM with which it is currently registered. Therefore, the Unified CM Group configuration of the phone's device pool provides redundancy for these services.

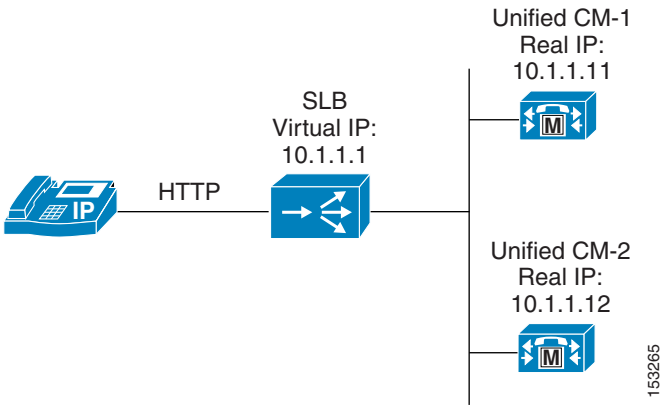
If Services Provisioning is set to External URL or both, while most of the back-end processing of a phone service occurs on a web server, the phones still depend upon Unified CM to inform them of the service URLs for their subscribed phone services. Given the architecture of IP phone service functionality and the message flows shown in Figure 18-1 and Figure 18-2, the following two main failure scenarios should be considered.

Failure Scenario 1: Server with Cisco CallManager Cisco IP Phone Services Fails

Redundancy in this case depends upon some type of server load balancing (SLB), as illustrated in Figure 18-3, where a virtual IP address (or DNS-resolvable hostname) is used to point to one or more Unified CM servers. This virtual IP address (or DNS-resolvable hostname) is used when configuring the URL Services parameter. The SLB device is configured with the real IP addresses of the Unified CM subscriber nodes. Thus, a Unified CM server failure does not prevent the IP Phone Services subscription list from being returned to the phone when the phone's Services or Applications button is pushed. In addition, phone services such as Extension Mobility and Unified CM Assistant that run on a Unified CM server are also potentially made redundant by this method. (See High Availability for Extension Mobility, page 18-15, and High Availability for Unified CM Assistant, page 18-24.)

Most SLB devices can be configured to monitor the status of multiple servers and automatically redirect requests during failure events.

Figure 18-3 Method for Providing Redundancy for Phone Services



Failure Scenario 2: External Web Server Hosting a Particular IP Phone Service Fails

In this scenario, the connection to the Unified CM server is preserved, but the link fails to the web server hosting the user-subscribed phone service. This is an easier scenario to provision for redundancy because the IP phone is still able to access the Unified CM server when the Services or Applications button is pressed. In this case, the IP phone is similar to any other HTTP client accessing a web server. As a result, you can again use some type of SLB functionality (similar to the one indicated in Figure 18-3) to redirect the HTTP request from the phone to one or more redundant web servers hosting the user-subscribed phone service.

Capacity Planning for IP Phone Services

Cisco Unified IP Phone Services act, for the most part, as an HTTP client. In most cases it uses Unified CM only as a redirect server to the location of the subscribed service. Because Unified CM acts as a redirect server to the phone service, there typically is minimal performance impact on Unified CM when a user initiates a phone service request by pressing the Services key, but a large number of requests (hundreds of requests per minute or more) could affect the server performance. To minimize the impact on the server performance, if an external URL does not need to be specified for the IP Phone Services, Cisco generally recommends leaving the Services Provisioning Enterprise Parameter set to **Internal**. If Services Provisioning has to be set to **External URL** or **Both**, or if you are using a large number phones that do not have the ability to retrieve the list of services from their configuration file (such as the Cisco

Unified IP Phone 7960), carefully select the node that will provide the Cisco Unified IP Phone Services list. For example, consider using the Unified CM TFTP servers instead of the Unified CM publisher if the load on the publisher is already high, or consider using Unified CM subscribers that are not handling a lot of traffic.

Note In the case of Extension Mobility and Unified CM Assistant phone service, Unified CM acts as more than a redirect server, and additional performance impacts should be considered. See the sections on Extension Mobility, page 18-7, and Unified CM Assistant, page 18-19, for specific performance and scalability considerations for these applications.

Because the IP Phone is either an HTTP client or server, estimating the required bandwidth used by an IP Phone service is similar to estimating the bandwidth of an HTTP browser accessing the same text as HTTP content residing on a web hosting server.

Design Considerations for IP Phone Services

With the exception of the integrated Extension Mobility and Unified CM Assistant applications' Phone Services, IP Phone services must reside on a separate off-cluster non-Unified CM web server. Running phone services other than Extension Mobility and Unified CM Assistant on the Unified CM server node is not supported.

Most Cisco IP Phones support content with text and graphics. Some phones such as the Cisco Unified IP Phone 7911G support only text-based XML applications. Some Cisco endpoints such as the Cisco TelePresence endpoints might not support Cisco IP Phone Services.

Extension Mobility

The Cisco Extension Mobility (EM) feature enables users to configure a Cisco Unified IP Phone as their own, on a temporary basis, by logging in to that phone. After a user logs in, the phone adopts the user's individual device profile information, including line numbers, speed dials, services links, and other user-specific properties of a phone. For example, when user X occupies a desk and logs in to the phone, that user's directory number(s), speed dials, and other properties appear on that phone; but when user Y uses the same desk at a different time, user Y's information appears. The EM feature dynamically configures a phone according to the authenticated user's device profile. The benefit of this application is that it allows users to be reached at their own extension on any phone within the Unified CM cluster, regardless of physical location, provided the phone supports EM.

This section examines the following design aspects of the Extension Mobility feature:

- Unified CM Services for Extension Mobility, page 18-8
- Extension Mobility Architecture, page 18-8
- Extension Mobility Security, page 18-13
- Extension Mobility Cross Cluster (EMCC), page 18-9
- High Availability for Extension Mobility, page 18-15
- Capacity Planning for Extension Mobility, page 18-17
- Design Considerations for Extension Mobility, page 18-18

Unified CM Services for Extension Mobility

The EM application relies on the Cisco Extension Mobility service, which is a feature service and which you must activate manually from the Serviceability page.

EM also relies on the Cisco Extension Mobility Application network service, which is activated automatically on all Unified CM nodes during installation.

The Cisco Extension Mobility Application service is a network service that provides an interface between the EM user phone and the Cisco Extension Mobility service. In addition, the Cisco Extension Mobility Application service subscribes to the change notification indications within the cluster and maintains a list of nodes in the cluster that have an active Cisco Extension Mobility service.

Extension Mobility Architecture

Figure 18-4 depicts the message flows and architecture of the EM application. When a phone user wants to access the EM application, the following sequence of events occurs:

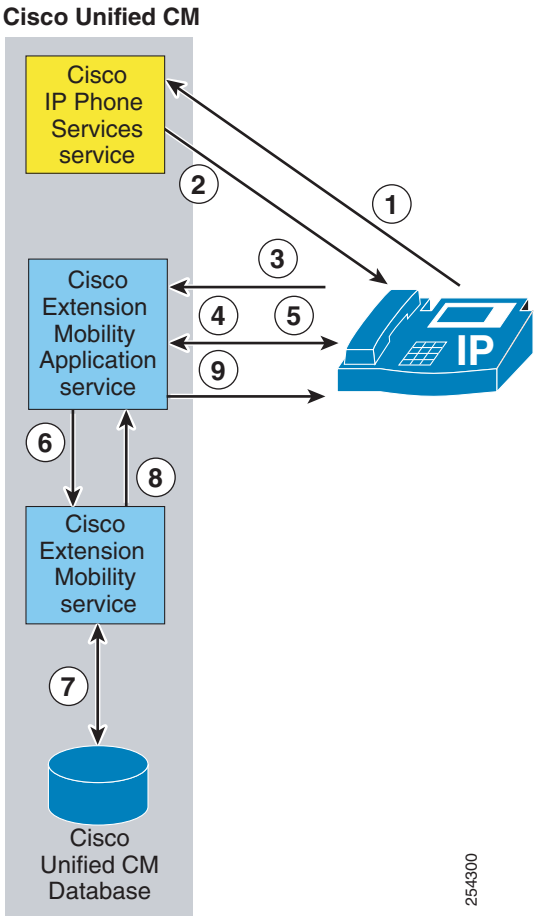
1. When the user presses the Services or Applications button on the phone, this action generates a call to the URL specified under the URL Services parameter on the Enterprise Parameter configuration page (see step 1 in Figure 18-4).
2. An HTTP/XML call is generated to the IP Phone Services, which returns a list of all services to which the user's phone is subscribed (see step 2 in Figure 18-4).

Note With the Services Provisioning enterprise parameter set to Internal, steps 1 and 2 are bypassed. Alternatively, with Services Provisioning set to External URL or Both, a Service URL button can be configured for EM on a user's phone so that the user can press a line or speed-dial button to generate a direct call to the Cisco Extension Mobility Application service, also bypassing steps 1 and 2.

3. Next the user selects the Extension Mobility phone service listing. This selection in turn generates an HTTP call to the Cisco Extension Mobility Application service, which serves as the interface between the phone and the Cisco Extension Mobility service (see step 3 in Figure 18-4).
4. The Cisco Extension Mobility Application service then forwards an XML response back to the phone requesting user login credentials (userID and PIN) or, if the user is already logged in, a response asking if the user wants to log off the phone (see step 4 in Figure 18-4).
5. Assuming the user is attempting to log in, the user must use the phone's keypad to enter a valid userID and PIN. After the user presses the Submit softkey, a response containing the userID and PIN just entered is forwarded back to the Cisco Extension Mobility Application service (see step 5 in Figure 18-4).
6. The Cisco Extension Mobility Application service next forwards this login information to the Cisco Extension Mobility service, which interacts with the Unified CM database to verify the user's credentials (see step 6 in Figure 18-4). The Cisco Extension Mobility Application service subscribes to cluster change notification, and it maintains a list of all nodes in the cluster with the Cisco Extension Mobility service activated. Therefore, in case the Cisco Extension Mobility service is not running on the same Unified CM node, the Cisco Extension Mobility Application service forwards the login information to other Unified CM nodes that are running the Cisco Extension Mobility service.

7. Upon successful verification of the user's credentials, the Cisco Extension Mobility service also interacts with the Unified CM database to read and select the appropriate user device profile and to write needed changes to the phone configuration based on this device profile (see step 7 in Figure 18-4).
8. Once these changes have been made, the Cisco Extension Mobility service sends back a successful response to the Cisco Extension Mobility Application service (see step 8 in Figure 18-4).
9. The Cisco Extension Mobility Application service, in turn, sends a reset message to the phone, and the phone resets and accepts the new phone configuration (see step 9 in Figure 18-4).

Figure 18-4 EM Application Architecture and Message Flow

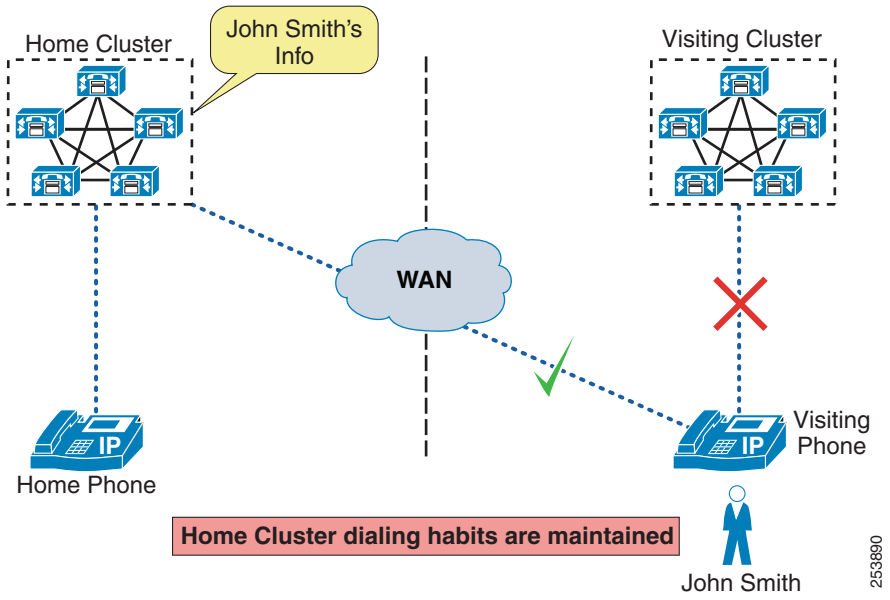


Extension Mobility Cross Cluster (EMCC)

Unified CM provides the ability to perform Extension Mobility logins between clusters within an enterprise with a new feature called Extension Mobility Cross Cluster (EMCC). It is important to understand the high-level architecture of EMCC. The EMCC feature employs the concepts of a home cluster and a visiting cluster, and these terms are defined from the perspective of the user performing the login. When a user travels to an office and attempts to log in to a phone, if the cluster to which this phone

is registered does not contain the user's information in its database, then this cluster is considered a visiting cluster and the phone is hereinafter referred to as the visiting phone. Figure 18-5 illustrates the concept of home and visiting clusters.

Figure 18-5 EMCC Home Cluster and Visiting Cluster



The EM service in the visiting cluster attempts to locate the home cluster of the user by sending out queries to each of the EMCC remote clusters that have been configured in Unified CM. When the user's home cluster responds positively, this initiates communications between the EM services of both clusters to exchange information that essentially brings the device information into the home cluster database and allows the home cluster to build a configuration file for this visiting phone. This configuration file incorporates some device configuration from the visiting cluster, configuration parameters from the home cluster, and the user's device profile in the home cluster. Once the home cluster TFTP server has a configuration file for this visiting phone, a reset issued by the visiting cluster forces the visiting phone to download a small configuration from the visiting cluster, which further instructs it to download certificates and a full configuration from the home cluster. Ultimately, the visiting phone cross-registers with the home cluster. This means that all call control signaling occurs between a home cluster Unified CM subscriber and the visiting phone, and the user's home cluster dialing habits are maintained.

For a step-by-step description of the EMCC login process, refer to the Extension Mobility Cross Cluster information in the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

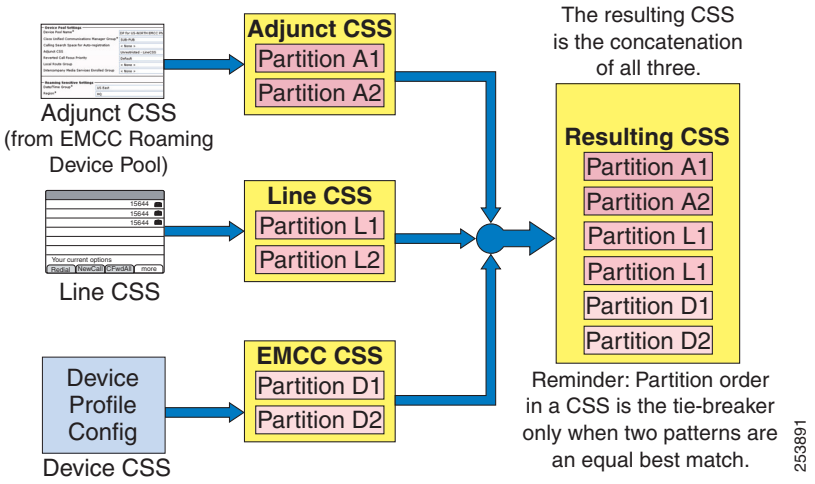
https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Call Processing

EMCC call processing behavior is also critical to understand because it impacts dial plan design. When a user has logged into a phone in a visiting cluster, any digits dialed by the user are analyzed by the home cluster according to the visiting phone's assembled call search space (CSS), which is a concatenation of the Adjunct CSS in the home cluster's device pool for the visiting phone (referred to as the EMCC

roaming device pool), the Line CSS configured on the directory number associated with the user's device profile, and the EMCC CSS configured on the user's device profile. Figure 18-6 illustrates the resulting CSS for an EMCC phone.

Figure 18-6 Resulting CSS for an EMCC Phone



The Adjunct Calling Search Space is a new call routing configuration parameter that is used by EMCC to intercept and route emergency numbers for users from a visiting cluster. The Adjunct CSS contains a partition with directory numbers such as 911, 112, or 999, that route the calls to the visiting cluster and allow the call to reach emergency services local to the physical phone's location. For more information on Adjunct Calling Search Spaces and the EMCC roaming device pool and how it is associated with a visiting phone, refer to the Extension Mobility Cross Cluster information in the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Note The EMCC roaming device pool associated with the EMCC feature is not related to the roaming device pool associated with the Device Mobility feature.

EMCC users must be aware that, when placing calls, they will be leveraging their home Unified CM routes and numbering plan. For example, if a user from Cluster A logs into a phone from Cluster B and wants to place a call to the directory number of a Cluster B phone located right next to it, the user would have to dial the appropriate pattern as if the user was placing the call from Cluster A to the phone in Cluster B. This implies that the home cluster may initiate an intercluster trunk call from Cluster A to Cluster B, but the media will flow locally between the visiting phone and the remote phone.

If the EMCC clusters have been deployed using +E.164 numbering, then the users should already be accustomed to dialing the full number of the target number and will not need to alter their dialing habits.

With PSTN routed calls, there are two different configurations that affect call processing behavior:

- Route patterns that do not use the Local Route Group (LRG) feature
- Route patterns that use the LRG feature

When an EMCC logged-in user dials a PSTN call, if the digit analysis matches a route pattern that ultimately leads to a voice gateway (either via the route list and route group construct or configured directly to a voice gateway), the call is offered out the gateway. If the Standard Local Route Group (Standard LRG) feature is not in use, and the call involves a voice gateway associated with the home cluster; therefore media will flow between the visiting phone (typically across a WAN) back to the voice gateway. When the route pattern leads to a route list configured to use Standard LRG, the behavior changes. (For more information about LRG, see Local Route Group, page 14-31.) When Unified CM logic must invoke a Standard LRG for an EMCC logged-in device, it recognizes the endpoint as an EMCC device and sends the PSTN call across a designated EMCC-specific SIP trunk to the visiting cluster to which this visiting phone is normally registered.

Note Only one SIP trunk with an EMCC trunk service type is required per cluster. There is no destination information configured on this trunk; that information is gathered dynamically when adding and updating an EMCC remote cluster.

When a call invite is received on the EMCC SIP trunk in the visiting cluster, the visiting cluster again performs digit analysis on the called number according to the CSS of the trunk (or alternatively, according to the CSS of the visiting phone's original device configuration), and routes the call accordingly. There is additional information included in a SIP invite across an EMCC SIP trunk, namely the device name of the visiting phone. This enables the visiting cluster to determine the configured device CSS of the visiting phone in the database (if required); and if the digit analysis results in matching a route pattern that ultimately points to the Standard LRG, the visiting cluster is able to determine the configured Standard LRG for this visiting phone. The Standard LRG in the visiting cluster will typically contain voice gateways associated with the visiting cluster, therefore the PSTN call is offered out a voice gateway local to the visiting phone.

The difference between LRG and non-LRG call processing behavior is critical when considering calls to emergency numbers. While the use of Local Route Groups (LRGs) is not required cluster-wide for an EMCC deployment, the EMCC logged-in phones must have access to an LRG in order to route emergency calls correctly. An LRG is required to correctly route an emergency call to a visiting cluster so that the call can be placed through an appropriate voice gateway local to the visiting phone. The Adjunct Calling Search Space in the roaming device pool configuration for an EMCC device enables an administrator to add emergency route patterns that will use an LRG for EMCC logged-in devices, but it will not affect emergency dialing for other devices in the home cluster. As discussed earlier, an EMCC logged-in phone will be associated with a device pool (by means of geolocations) that represents all phone devices from another cluster. The device pool's Adjunct Calling Search Space allows for the visiting cluster's emergency route pattern to be configured so that only emergency calls for an EMCC logged-in phone will be sent through an LRG. So even if the home and visiting clusters use the same emergency route pattern, the EMCC logged-in phone's emergency call will route through the LRG to the visiting cluster. Once the call is received at the visiting cluster through the EMCC SIP trunk, the visiting cluster dial plan will be responsible for further processing of the call.

Note If any cluster supporting EMCC is also using Cisco Emergency Responder for emergency call processing, refer to the latest version of the *Cisco Emergency Responder Administration Guide* for information on how to configure the dial plan to support the deployment, available at https://www.cisco.com/c/en/us/support/unified-communications/emergency-responder/products-maintenance-guides-list.html.

Note If Standard LRGs are already deployed for the emergency route pattern, and if the home and visiting clusters use the same emergency dial string, use of the Adjunct CSS is not required.

For detailed EMCC call processing examples and configuration, refer to the Extension Mobility Cross Cluster information in the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Media Resources

All media resources except for RSVP agents are allocated from the home cluster according to the media resource group list of the device pool assigned to the visiting phone. Conferencing, transcoding, and music on hold all function as normal, with the difference being that media is streaming between the visiting phone and media resources across (typically) a WAN separating the home and visiting clusters. When an EMCC logged-in user makes a call that requires use of an RSVP agent, the Unified CM EMCC logic is able to determine it is a visiting phone, and it sends a resource request across the EMCC SIP trunk to the remote cluster to which the visiting phone belongs. The device name of the visiting phone is included in this request, which enables the visiting cluster to verify the RSVP agent media resources that are normally assigned to this visiting phone and to allocate its use for the call.

Extension Mobility Security

Unified CM provides the ability to create an Extension Mobility secure service URL using HTTPS. This encrypts the entire EM login/logout exchange. Cisco recommends configuring a secure service URL for Extension Mobility. If there are phones deployed for EM that do not support HTTPS, a non-secure service URL must also be configured. When secure and non-secure service URLs exist for the service, phones that support HTTPS use the secure service URL by default. For a complete list of phones that support HTTPS, refer to the HTTPS information in the latest version of the *Security Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

The EM feature provides an optional level of security for EM login and logout requests by validating the source IP address of the request. By default, EM does not perform this request validation; therefore, to enable EM security, the administrator must set the cluster-wide service parameter Validate IP Address to true.

For organizations that implement a web proxy to handle EM login and logout HTTP requests, the Allow Proxy service parameter must be set to true. A proxy server, while forwarding the HTTP request, will set the via-field of the HTTP header with its hostname. If there are multiple proxy servers between the device and Unified CM, and if the request is forwarded by all the servers, then the via-field in the HTTP header will have a comma-separated list of hostnames for each of the proxy servers in the forwarding path. The Allow Proxy service parameter, if set to true, will allow EM login and logouts received via a web proxy. In addition, if the proxied EM requests use the source IP address of the proxy server, this IP address must also be configured in the Trusted List of IPs service parameter.

With support for HTTPS and Security By Default starting in Unified CM 8.*x*, and with the introduction of secure phones support for EMCC in Unified CM 9.*x*, the intercluster interactions of EMCC require some extra steps to ensure that clusters can communicate with each other in a secure manner. In particular, all clusters that participate in EMCC must export their Tomcat (web) and TFTP certificates to a central sFTP server. Exporting the CAPF certificates is also required if phones used for EMCC will be in secure mode. These security certificates are all combined, and then each cluster must import the combined certificate into its cluster. It is important to remember that any time a new node that may participate in EMCC is added to the cluster, or if a certificate on any existing node is updated, the process

of exporting, combining, and importing must be repeated. All of these steps have been streamlined via Unified CM Serviceability administration. For details on EMCC configuration, refer to the Extension Mobility Cross Cluster information in the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Support for Phones in Secure Mode

Starting with Cisco Unified CM 9.*x*, users can log in through EMCC using phones in secure mode — that is, phones with an authenticated or encrypted Device Security Profile. When a user logs in on a phone in secure mode, the configuration in the device security profile (such as the device security mode, TFTP encrypted option, and transport protocol) is transferred to the home cluster, allowing the phone to operate in the same secure mode as it was originally in the visiting cluster. For example, if the phone is configured with the encrypted device security mode in the visiting cluster and the user logs in through EMCC, the phone still operates in the encrypted device security mode with a secure TLS channel for signaling and sRTP for media. However, one condition is that the home cluster security mode must be configured as mixed mode. If the home cluster is configured as non-secure instead, the EMCC login will fail. If the phone is not in secure mode, the phone continues to operate in a non-secure mode after the EMCC login, regardless of whether the visiting cluster is in mixed mode or non-secure mode. Table 18-2 indicates this behavior.

Unified CM 8.*x* supports EMCC but not with phones in secure mode. For this reason, EMCC login attempts from a phone in secure mode registered to a visiting cluster running Unified CM 8.*x* will fail, regardless of whether the home cluster is running Unified CM 8.*x* or a later release. Similarly, EMCC login attempts from a phone in secure mode to a home cluster running Unified CM 8.*x* will fail, regardless of whether the visiting cluster is running Unified CM 8.*x* or a later release. Table 18-2 indicates this behavior.

Table 18-2 Phone Security Mode After EMCC Login

Visiting Cluster	Home Cluster Running Unified CM 8.*x*	Home Cluster Running Unified CM 9.*x* or Later Release	
	Mixed Mode or Non-Secure Mode	Mixed Mode	Non-Secure Mode
Phone in secure mode; visiting cluster running Unified CM 8.*x*	EMCC login fails	EMCC login fails	EMCC login fails
Phone in secure mode; visiting cluster running Unified CM 9.*x* or later release	EMCC login fails	Secure mode	EMCC login fails
Phone in non-secure mode; visiting cluster running Unified CM 8.*x* or later release (Visiting cluster in mixed mode or non-secure mode)	Non-secure mode	Non-secure mode	Non-secure mode

Note As of Cisco Unified CM 9.0, the EMCC SIP trunk cannot be configured with a secure profile. Therefore, calls to the local PSTN do not use a secure channel for signaling. However, the media is encrypted if the phone and PSTN gateway are configured in a secure mode.

High Availability for Extension Mobility

According to the EM architecture illustrated in Figure 18-4, reads and writes to the Unified CM database are required. EM is a user-facing feature, and database writes pertaining to EM can be performed by subscriber nodes. Therefore, if the Unified CM publisher is unavailable, EM logins and logouts are still possible.

From a redundancy perspective, the following component levels of redundancy must be considered for full EM resiliency:

- Cisco CallManager Cisco IP Phone Services

  High availability for the CallManager Cisco IP Phone Services is obtained by using the Services Provisioning service parameter or by using a load balancer device pointing to multiple Unified CM nodes running the Cisco CallManager Cisco IP Phone Services. For more details, see High Availability for IP Phone Services, page 18-5.

- Cisco Extension Mobility service

  High availability for the Cisco Extension Mobility service is obtained by activating the Cisco Extension Mobility service on multiple Unified CM nodes.

  Note While the Cisco Extension Mobility service can be activated on more than two nodes, a maximum of two nodes should actively handle login/logout requests at any given time. If a load balancer is used, configure the load balancer to send Extension Mobility requests to only two Unified CM nodes. The load balancer should start sending login/logout requests to other nodes running the Cisco Extension Mobility service only in case of failure.

Cisco recommends deploying a server load balancer device to load-balance the requests across two Unified CM nodes and to provide redundancy. Without a server load balancer, load balancing would be uneven and the redundancy would be manual. For example, two EM IP Phone services could be configured on each phone. If one Unified CM node is not reachable, the end user would have to manually select the other EM IP Phone service to reach the other node.

Note While it is possible to provide redundancy for the EM IP Phone service by relying on end users to manually select an EM IP Phone service from a list of EM IP Phone services, achieving high availability in this manner can be problematic. Because there is no control over which EM IP Phone service a user might select from the phone services menu (or assigned feature keys), there is no way to ensure that the EM login/logout load is balanced between Unified CM nodes handling EM login/logout requests. Further, end user behavior when encountering delay in response of the EM service, which is typical in a failure scenario, will usually exacerbate the situation as users cancel EM service calls and select alternate EM IP Phone service. This can lead to added congestion and load on the network as well as on the remaining Unified CM node handling EM login/logout requests.

A deployment with two Unified CM nodes running the Cisco Extension Mobility service provides the highest capacity in terms of number of login/logout requests per minute. (See Capacity Planning for Extension Mobility, page 18-17, for details.) It also provides redundancy. However, in case of failure, the login/logout request capacity is reduced because there is only one node left. Therefore, to achieve the highest login/logout capacity and maintain this capacity in case of failure, the Cisco Extension Mobility service should be activated on additional Unified CM nodes. A load balancer should be deployed and configured to send Extension Mobility requests to only two Unified CM nodes at a given

time. If one Unified CM node fails, the load balancer can start sending Extension Mobility login/logout requests to another Unified CM node so that two Unified CM nodes are still processing the Extension Mobility requests. This would allow the Extension Mobility capacity to be maintained.

Note Cisco does not recommend a redundancy design using DNS A or SRV records with multiple IP listings. With multiple IP addresses returned to a DNS request, the phones must wait for a timeout period before trying the next IP address in the list, and in most cases this results in unacceptable delays to the end user. In addition, this can result in more than two subscriber nodes with the Cisco Extension Mobility Application service enabled to handle login/logout requests, which is not supported.

With EMCC, remote clusters are administratively added via Unified CM web administration by specifying a single FQDN or IP address of a Unified CM subscriber node running the EM service in the remote cluster. The EM services between the two clusters provide information about the Unified CM version, an ordered list of EM Service nodes for EMCC EM Service communications, which EMCC SIP trunk services are enabled (PSTN Access and/or RSVP Agent) in the remote cluster, and an ordered list of up to three remote Unified CM nodes that handle EMCC SIP trunk operations for each EMCC service. EMCC EM service communications over HTTPS include locating users' home clusters, exchanging information during EMCC logins, and remote cluster updates. Upon an initial update, a remote cluster's Extension Mobility Application service is queried, which will return the first three EM Service nodes in its list. This ordered list determines which remote cluster EM Service nodes will be used for EMCC communications.

The remote cluster obtains the information regarding primary, secondary, and tertiary options for EMCC PSTN Access and RSVP Agent services from the Unified CM Group that is associated with the device pool of the assigned EMCC SIP trunk for those services. This ensures that, if the primary Unified CM subscriber handling the EMCC SIP trunk is offline, then the EMCC SIP trunk call will be handled by the secondary Unified CM subscriber, and so on.

Once a phone is logged in through EMCC, redundancy is provided for the phone in the form of the Unified CM Group configured in its assigned EMCC device pool. If the visiting phone is located in a remote site and there is a WAN outage in which both the visiting and home cluster are unreachable, then the SRST reference from the visiting cluster is maintained by the EMCC phone. Therefore, an EMCC logged-in phone will still be able to register with the appropriate SRST router in the site where it is located. The EMCC logged-in user's DID most likely will not be associated with the local gateway(s) at the SRST site, so incoming calls will still be routed based on the call forwarding rules on the user's home cluster. While in SRST mode, the user will also have to adapt to the visiting SRST site's configured dial habits during SRST failover registration. For additional examples of an EMCC logged-in phone's behavior during a networking failure, refer to the Cisco Extension Mobility Cross Cluster section in the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Cisco also recommends configuring a default and backup Unified CM TFTP server to be used for visiting phones to download EMCC configuration files that will allow them to register with the home cluster. This is configured under EMCC Feature Configuration.

Capacity Planning for Extension Mobility

With a single Unified CM running the Cisco Extension Mobility application, the maximum cluster-wide capacity is 250 logins and/or logouts per minute when the Unified CM node is deployed with the 7,500-user or 10,000-user VM configuration. Cisco Extension Mobility login and logout functionality can be distributed across a pair of subscriber nodes to increase login/logout cluster capacity. A load balancer device can be used, or to manually distribute the EM load evenly between the two subscriber nodes, the phones should be divided into two groups, with one group of phones subscribed to an EM phone service pointing to one of the subscriber nodes and the other group of phones subscribed to a second EM phone service that is pointing to a second subscriber node. When the EM load is distributed in this way, evenly between Unified CM nodes using the 7,500-user or 10,000-user VM configuration, the maximum cluster-wide capacity is 375 sequential logins and/or logouts per minute.

Note The Cisco Extension Mobility service can be activated on more than two nodes for redundancy purposes, but Cisco supports a maximum of two subscriber nodes actively handling logins/logouts at any given time.

Note Enabling EM Security does not diminish performance.

The EMCC login/logout process requires more processing resources than intracluster EM login/logout, therefore the maximum supported login/logout rates are lower. In the absence of any intracluster EM logins/logouts, Unified CM supports a maximum rate of 75 EMCC logins/logouts per minute when using the 7,500-user or 10,000-user VM configuration. Most deployments will have a combination of intracluster and intercluster logins/logouts occurring. For this more common scenario, the mix of EMCC logins/logouts (whether acting as home cluster or visiting cluster) should be modeled for 40 per minute while the intracluster EM logins should modeled for 185 logins/logouts when using a single EM login server. The intracluster EM login rate can be increased to 280 logins/logouts per minute when deploying two Unified CM nodes in dual EM service configuration and using the 7,500-user or 10,000-user VM configuration.

For more details on the capacity limits, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

EMCC logged-in devices (visiting phones) consume twice as many resources as any other endpoint in a cluster. The maximum supported number of EMCC logged-in devices is 2,500 per cluster, but this also decreases the theoretical maximum number of other devices per cluster from 30,000 to 25,000. Even if the number of other registered devices in the cluster is reduced, the maximum supported number of EMCC logged-in devices is still 2,500.

There is no technical limit to the number of EMCC remote clusters that can be added to a cluster; however, the full-mesh requirement will increase the load on the EM service as the number of remote clusters increases. For a high number of sites (more than 10), the EM CPU should be monitored by means of the Cisco Real-Time Monitoring Tool (RTMT).

Design Considerations for Extension Mobility

The following guidelines and restrictions apply with regard to the deployment and operation of EM within the Unified CM environment:

- EM users should not move between locations or sites within a cluster when Automated Alternate Routing (AAR) and/or the Voice over PSTN (VoPSTN) deployment model are in use.

  EM functionality relies on the use of the IP network for routing calls. Call routing via the PSTN is more problematic because E.164 PSTN numbers are static and the PSTN is unable to account for movement of EM user directory numbers (DNs) from their home sites. AAR relies on the PSTN for call routing, as does the VoPSTN deployment model. In both cases, EM user movement between locations and sites is supported only if all sites the user is traversing are in the same AAR group. For additional information, see Extension Mobility, page 14-84.

- Restarting the Cisco Extension Mobility service or the node on which the service is running will affect auto-logout settings.

  If the Cisco Extension Mobility service is stopped or restarted, the system does not auto-logout users who are already logged in after the expiration of the maximum login interval. These phones will either have to be logged out manually or wait until the daily database clean-up process runs (typically at midnight).

Some Cisco endpoints such as the Cisco TelePresence endpoints might not support Extension Mobility.

WebDialer supports the use of phones logged in using Extension Mobility. For more information, please see WebDialer, page 18-34.

Design Considerations for Extension Mobility Cross Cluster (EMCC)

The following design considerations apply when deploying EMCC.

General Design Considerations

- Prior to Unified CM release 9.1(1), EMCC requires that all users must be unique across *all* clusters in the enterprise. If LDAP synchronization is maintaining common users for multiple clusters, some type of filtering must be applied.
- Starting with Unified CM release 9.1(1), the same user ID can exist in multiple cluster; however, only one cluster should be defined as a **home cluster** for the user. When a user attempts to log in on a cluster that has the **home cluster** option selected for the user, the cluster will perform a local EM login and will not attempt an EMCC login with the remote cluster(s).
- Consider the network delay between clusters in combination with the features you plan to use. As the visiting phone is registered with the home cluster, features will work. However, depending on the network delay for a given deployment, all applications and features might not meet user requirements. Testing might be required to determine the usability of features for a given network.

  For example, EMCC supports dynamic CTI control of a visiting phone. But if an offhook is issued via an application and it takes 1 second before the phone goes offhook, this might be acceptable for an office worker but might not be acceptable for a call center agent.
- Phone load firmware is not enforced during the login process. Instead, the visiting cluster phone load information is maintained so that cross-registration does not result in new phone firmware downloads.

- If the home cluster locale is different than that of the visiting cluster, the phone will download the new locale from the visiting cluster TFTP server. If it is not available, then the phone will not change locales and instead will maintain the visiting cluster locale.
- The total number of EMCC logins is controlled by the total number of EMCC inserted devices in the Bulk Administration Tool (BAT).
- EMCC supports RSVP-based and Unified CM locations-based call admission control.
- Except for RSVP agents, all other media resources are allocated from the home cluster according to the media resource group list associated with the EMCC roaming device pool.
- Audio and video codecs are determined by the EMCC region settings. These settings override normal region configuration for EMCC registered phones. All EMCC region parameters must be configured with the same values in all clusters. If they are different, RSVP Agent for that cluster will be disabled by the remote cluster update operation.
- For the EMCC roaming device pool to be assigned correctly, EMCC-capable phones must have a geo-location configured via device configuration or a device pool.

Call Processing Design Considerations

- Incoming calls for a user's directory number will always be received on a home cluster voice gateway, therefore RTP media will flow between the visiting phone and the home gateway for incoming calls.
- Calls sent across the EMCC SIP trunk will have gone through digit manipulation in the home cluster. The called number may require manipulation to match visiting cluster route patterns.
- Verify configured codec capabilities of H.323 and SIP gateways in the home cluster. For example, if home cluster gateways are configured to accept only G.711 calls and the EMCC region bandwidth is set to 8 kbps (G.729), a transcoder is required to complete the call. Alternatively, the H.323 or SIP gateway dial peers may be configured to allow for G.729 in addition to G.711.
- Design considerations must be made regarding the calling party for EMCC emergency calls. Depending on dial plan configurations, the calling party number leaving the visiting cluster gateway may be the user's DID that is normally associated with the home cluster. This would require transforming the calling number incoming on the EMCC SIP trunk, on route patterns, or egressing on the visiting gateways.
- When EMCC is deployed with Cisco Emergency Responder, Emergency Responder should be deployed in all clusters handled by a single Emergency Responder cluster. If the visiting cluster is deployed with Emergency Responder and the home cluster is not, Emergency Responder will not be able to identify the visiting phone when the call arrives back to the visiting cluster.

Unified CM Assistant

Cisco Unified Communications Manager Assistant (Unified CM Assistant) is a Unified CM integrated application that enables assistants to handle incoming calls on behalf of one or more managers. With the use of the Unified CM Assistant Console desktop application or the Unified CM Assistant Console phone service on the assistant phone, assistants can quickly determine a manager's status and determine what to do with a call. Assistants can manipulate calls using their phone's softkeys and service menus or via the PC interface with either keyboard shortcuts, drop-down menus, or by dragging and dropping calls to the managers' proxy lines.

This section examines the following design aspects of the Unified CM Assistant feature:

- Unified CM Assistant Architecture, page 18-20
- High Availability for Unified CM Assistant, page 18-24
- Capacity Planning for Unified CM Assistant, page 18-26
- Design Considerations for Unified CM Assistant, page 18-28
- Unified CM Assistant Console, page 18-32

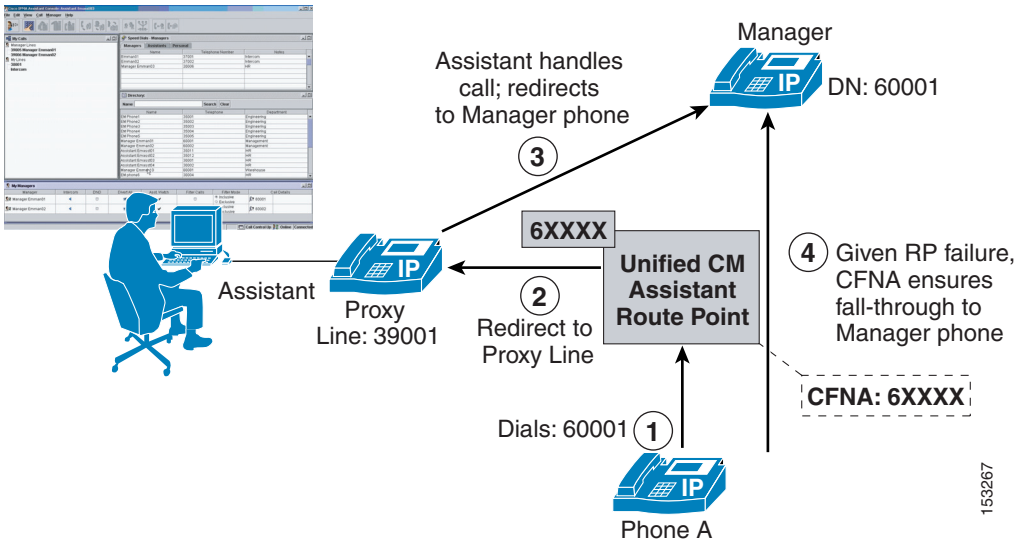
Unified CM Assistant Architecture

The Unified CM Assistant application can operate in two modes: proxy line mode and shared line mode. The operation and functionality of each mode is different, and each has specific advantages and disadvantages. Both modes can be configured within a single cluster. However, mixing modes on the same assistant is not allowed. A single assistant providing support for one or more managers can support those managers in either shared line mode or proxy line mode.

Unified CM Assistant Proxy Line Mode

Figure 18-7 illustrates a simple call flow with Unified CM Assistant in proxy line mode. In this example, Phone A calls the Manager phone with directory number (DN) 60001 (step 1). The CTI/Unified CM Assistant Route Point (RP) intercepts this call based on a configured DN of 6XXXX. Next, based on the Manager DN, the call is redirected by the route point to the Manager's proxy line (DN: 39001) on the Assistant's phone (step 2). The Assistant can then answer or handle the call and, if appropriate, redirect the call to the Manager's phone (step 3). In the event of Unified CM Assistant application failure or if the Unified CM Assistant RP fails, a fall-through mechanism exists via the Call Forward No Answer (CFNA) 6XXXX configuration of the RP, so that calls to the Manager's DN will fall-through directly to the Manager's phone (step 4).

Figure 18-7 Unified CM Assistant Proxy Line Mode

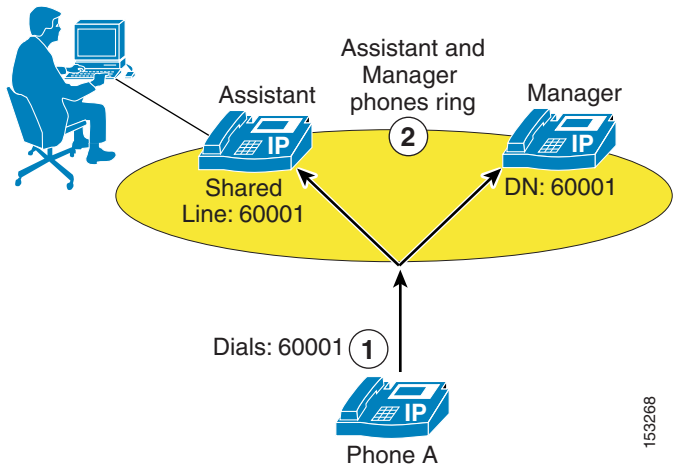


Note The CFNA fall-through mechanism illustrated in Figure 18-7 requires configuration of the same summarized digit-string as the Unified CM Assistant RP directory number in both the Forward No Answer Internal and Forward No Answer External fields under the Unified CM Assistant RP directory number configuration page. In addition, the calling search space (CSS) field for each of these call forward parameters should be configured with the calling search space containing the partition with which the Manager phone DNs are configured, so that the Manager phone DNs can be reached if the Unified CM Assistant RP or Unified CM Assistant application fails.

Unified CM Assistant Share Lined Mode

Figure 18-8 illustrates a simple call flow with Unified CM Assistant in shared line mode. In this example, Phone A calls the Manager phone with directory number (DN) 60001, which is a shared line on the Assistant phone (step 1). The call will ring at both the Assistant and Manager phones unless the Manager has invoked the Do Not Disturb (DND) feature, in which case the Assistant's phone will be the only phone that rings audibly (step 2).

Figure 18-8 Unified CM Assistant Shared Line Mode



In Unified CM Assistant shared line mode, the Unified CM Assistant RP is not needed or required for intercepting calls to the Manager phone. However, the Do Not Disturb (DND) feature on the Manager phone and the Unified CM Assistant Console desktop application still depend on the Cisco IP Manager Assistant (IPMA) and Cisco CTIManager services. Furthermore, in Unified CM Assistant shared line mode, features such as call filtering, call intercept, assistant selection, and Assistant Watch are not available.

Unified CM Assistant Architecture

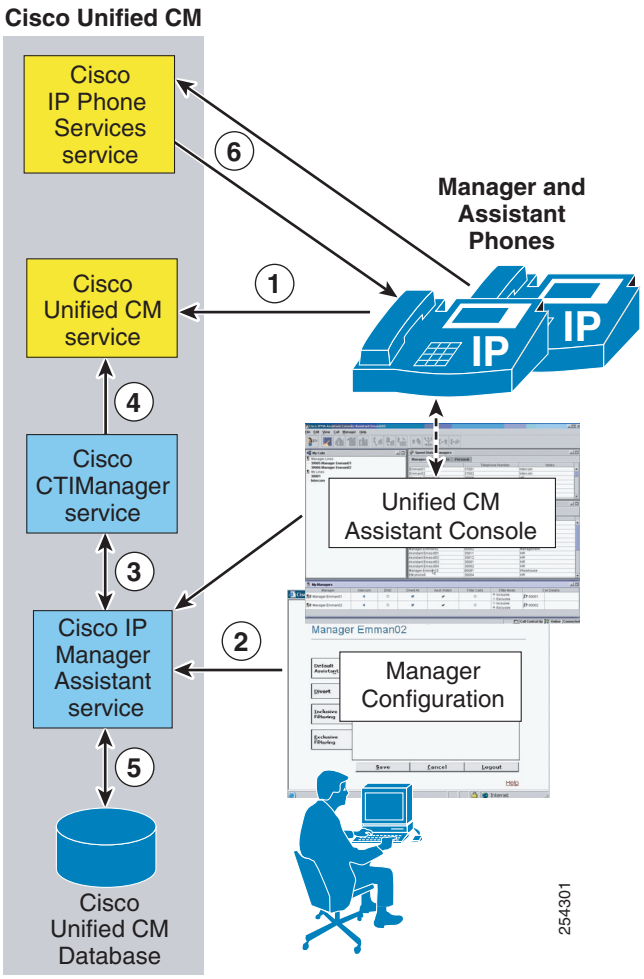
The architecture of the Unified CM Assistant application is as important to understand as its functionality. Figure 18-9 depicts the message flows and architecture of Unified CM Assistant. When Unified CM Assistant has been configured for Unified CM Assistant Manager and Assistant users, the following sequence of interactions and events can occur:

1. Manager and Assistant phones register with the Cisco CallManager Service, and the phone's keypad and softkeys are used to handle call flows (see step 1 in Figure 18-9).
2. Both the Unified CM Assistant Console desktop application and the Manager Configuration web-based application communicate and interface with the Cisco IP Manager Assistant service (see step 2 in Figure 18-9).
3. The Cisco IP Manager Assistant service in turn interacts with the CTIManager service for exchanging line monitoring and phone control information (see step 3 in Figure 18-9).
4. The CTIManager service passes Unified CM Assistant phone control information to the Cisco CallManager service and also controls the Unified CM Assistant RP (see step 4 in Figure 18-9).
5. In parallel, the Cisco IP Manager Assistant service reads and writes Unified CM Assistant application information to and from the Unified CM database (see step 5 in Figure 18-9).
6. The Manager may choose to invoke the Unified CM Assistant phone service by pushing the Services or Applications button, thus generating a call to the IP Phone Services service that will return a list of all services (including the Unified CM Assistant phone service) to which the phone is subscribed (see step 6 in Figure 18-9).

The Unified CM Assistant phone service is controlled by the Cisco IP Manager Assistant service, and configuration changes made by the Manager using the phone are handled and propagated via the Cisco IP Manager Assistant service.

Note With the Services Provisioning enterprise parameter set to Internal, steps 1 and 2 are bypassed. Alternatively, with Services Provisioning set to External URL or Both, a Service URL button can be configured for the Unified CM Assistant phone service on a user's phone so that the user can press a line or speed-dial button to generate a direct call to the Cisco IP Manager Assistant service, also bypassing steps 1 and 2.

Figure 18-9 Unified CM Assistant Architecture



Note While Figure 18-9 shows the IP Phone Services, Cisco CallManager, CTIManager, and Cisco IP Manager Assistant services all running on the same node, this configuration is not a requirement. These services can be distributed between multiple nodes in the cluster but have been shown on the same node here for ease of explanation.

High Availability for Unified CM Assistant

Unified CM Assistant application redundancy can be provided at two levels:

- Redundancy at the component and service level

  At this level, redundancy must be considered with regard to Unified CM Assistant service or server redundancy and CTIManager service redundancy. Likewise, the lack of publisher redundancy and the impact of this component failing should also be considered.

- Redundancy at the device and reachability level

  At this level, redundancy should be considered as it relates to Assistant and Manager phones, the Unified CM Assistant route point, and the Unified CM Assistant Console desktop application and phone service, as well as redundancy in terms of Assistant and Manager reachability.

Service and Component Redundancy

As shown in Figure 18-9, Unified CM Assistant functionality is primarily dependent on the Cisco IP Manager Assistant service and the Cisco CTIManager service. In both cases, redundancy is automatically built-in using a primary and backup mechanism. Up to three pairs of active and backup Unified CM Assistant servers (nodes running the Cisco IP Manager service) can be defined, for a total of six Unified CM Assistant servers within a single cluster. Active and backup Unified CM Assistant server pairs are configured using the Cisco IPMA Server IP Address, Pool 2 Cisco IPMA Server IP Address, and Pool 3 Cisco IPMA Server IP Address service parameters. With the configuration of these parameters, the required Cisco IP Manager service is made redundant. Given a failure of any of the primary Unified CM Assistant servers, the backup or standby Unified CM Assistant servers are able to handle Unified CM Assistant service requests. For each pair of Unified CM Assistant servers, only one Unified CM Assistant server can be active and handling request at a given time, while the other Unified CM Assistant server will be in a standby state and will not handle requests unless the active server fails.

In addition, two CTIManager servers or services can be defined for each Unified CM Assistant server using the CTIManager (Primary) IP Address and CTIManager (Backup) IP Address service parameters. By configuring these parameters, you can make the CTIManager service redundant. Thus, given a failure of a primary CTIManager, CTIManager services can still be provided by the backup CTIManager. If all Cisco IP Manager Assistant and CTIManager services on cluster nodes fail, the Unified CM Assistant route point, Unified CM Assistant Console desktop application and phone service, and in turn the Unified CM Assistant application as a whole will fail. However as noted previously, given a failure of the Unified CM Assistant application, the CFNA fall-through mechanism will continue to work, allowing calls to a Manager to be routed directly to the Manager's phone.

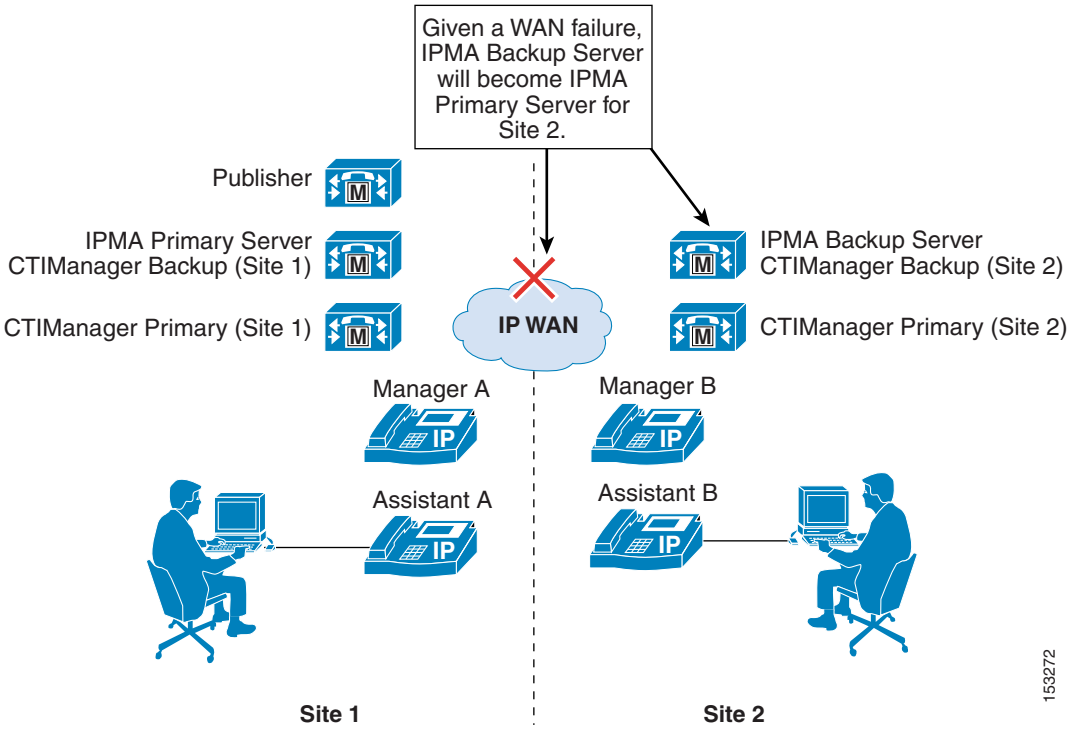
Note If configured in Unified CM Assistant shared-line mode, a complete failure of Cisco IP Manager Assistant and CTIManager service will not keep the Assistant from continuing to handle calls on behalf of the Manager because the phones will continue to shared a line. However, the Unified CM Assistant Console desktop application and phone service and the DND feature will not be available.

Figure 18-10 shows an example redundancy configuration for Unified CM Assistant and CTIManager primary and backup servers in a two-site deployment with clustering over the WAN. In order to provide maximum redundancy, a node at Site 1 is configured as the primary Unified CM Assistant server and a node at Site 2 is configured as the backup Unified CM Assistant server. In the event of a WAN failure, the backup Unified CM Assistant server at Site 2 will become a primary Unified CM Assistant server because the existing primary Unified CM Assistant server will be unreachable from Site 2. In this way, Unified CM Assistant servers can be made redundant in the clustering-over-the-WAN environment

given a WAN failure. Furthermore, with a primary and backup CTIManager configured at both Site 1 and Site 2, CTIManager is made redundant given a WAN failure, and additional redundancy is provided for a CTIManager failure at each site.

Note The redundancy scenario depicted in Figure 18-10 shows a special circumstance. During normal operation it is not possible to have any pair of Unified CM Assistant servers active at the same time. If an active and backup pair of Unified CM Assistant servers can communicate over the network, then one server will be in backup mode and cannot handle requests.

Figure 18-10 Unified CM Assistant Redundancy with Two-Site Clustering over the WAN



As previously mentioned, the publisher is a single point of failure when it comes to writing Unified CM Assistant information to the Unified CM database. Given a publisher failure, all aspects of the Unified CM Assistant application will continue to work; however, no changes to the Unified CM Assistant application configuration can be made. Configuration changes via the Unified CM Assistant Console desktop application, the Manager configuration web-based application, the phone softkeys, or the Unified CM Assistant phone service, will not be possible until the publisher is restored. This condition includes enabling or disabling features such as Do Not Disturb, DivertAll, Assistant Watch, and call filtering, as well as changing call filter and assistant selection configuration.

Device and Reachability Redundancy

Redundancy for Unified CM Assistant at the devices level relies on a number of mechanisms. First and foremost, manager and assistant phones as well as the Unified CM Assistant RP rely on the built-in redundancy provided by a combination of the device pool and Unified CM group configuration for device registration.

In addition, some devices rely on component services for additional redundancy and functionality. For example, the Unified CM Assistant RP also relies on CTIManager for call control functionality and therefore must rely on the primary and back CTIManager mechanism described in the previous section. The Unified CM Assistant Console desktop application also relies on the component services for redundancy and functionality. The Assistant Console desktop application supports automatic failover from the primary to the backup Unified CM Assistant server (and vice versa) in order to continue to handle incoming calls for managers. The amount of time this automatic failover will take can be controlled using the Cisco IPMA Assistant Console Heartbeat Interval and the Cisco IPMA Assistant Console Request Timeout service parameters. Although the heartbeat or keep-alive frequency can be configured so that failures of the Unified CM Assistant server are detected by the desktop application more quickly, be careful not to affect the network adversely by sending keep-alives too frequently. This consideration is especially important if there are a large number of Assistant Console desktop applications in use.

The Unified CM Assistant Console phone service, unlike the Unified CM Assistant Console desktop application, requires manual intervention for redundancy given the failure of the primary Unified CM Assistant server. If the primary Unified CM Assistant server goes down, assistants using the phone console will not see an indication of this condition. However, the assistant phone will receive a "Host not found Exception" message upon trying to use a softkey. In order to continue using the phone console with the backup Unified CM Assistant server, the user must manually select the secondary Unified CM Assistant phone service from the IP Services menu and log in again.

There are several other failover mechanisms which ensure that Manager and Assistant reachability are redundant. First, calls sent to a Manager's Assistant via the Unified CM Assistant application (in proxy line mode) can be forwarded to the Manager's next available Assistant if the call is not answered after a configured amount of time. If the next Assistant does not answer the call after the configured amount of time, the call can again be forwarded to the Manager's next available Assistant, and so on. The mechanism is configured using the Cisco IPMA RNA Forward Calls and Cisco IPMA RNA Timeout service parameters. Second, as mentioned previously, if all Cisco IP Manager Assistant and CTI services on cluster nodes fail, the Unified CM Assistant RP will become unavailable. However, based on the CFNA configuration of the Unified CM Assistant RP, calls to all Manager DNs will fall-through directly to the Manager phones so that Manager reachability is sufficiently redundant.

Capacity Planning for Unified CM Assistant

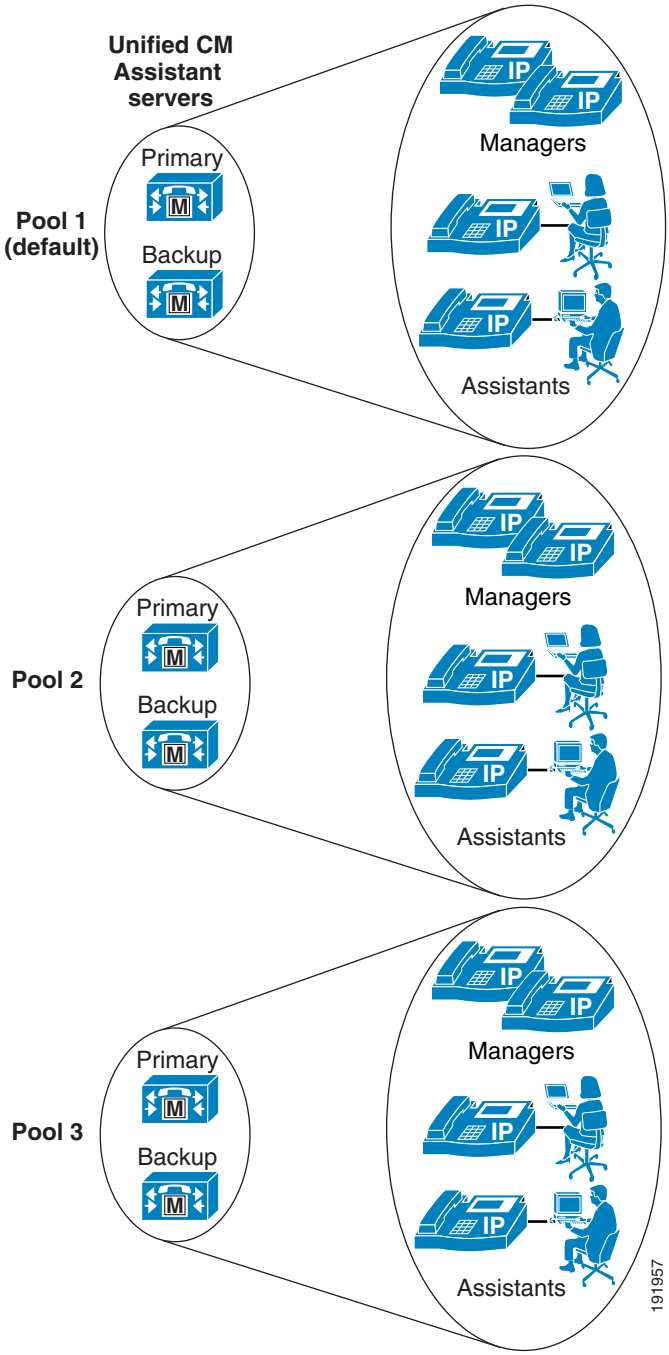
The Cisco Unified CM Assistant application supports the following capacities:

- A maximum of 10 Assistants can be configured per Manager.
- A maximum of 33 Managers can be configured for a single Assistant (if each Manager has one Unified CM Assistant-controlled line).
- A maximum of 3500 Assistants and 3500 Managers (7000 total users) can be configured per cluster using the 7,500-user or 10,000-user VM configuration.
- A maximum of three pairs of primary and backup Unified CM Assistant servers can be deployed per cluster if the Enable Multiple Active Mode advanced service parameter is set to True and a second and third pool of Unified CM Assistant servers are configured.

In order to achieve the maximum Unified CM Assistant user capacity of 3500 Managers and 3500 Assistants (7000 users total), multiple Unified CM Assistant server pools must be defined. As illustrated in Figure 18-11, up to three pools can be configured. Each pool consists of a primary and backup Unified CM Assistant server and a group of Managers and Assistants. Pool 1's Unified CM Assistant servers are configured with the Cisco IPMA Server (Primary/Backup) IP Address service parameters,

Pool 2's servers are configured with the Pool2: Cisco IPMA Server (Primary/Backup) IP Address advanced service parameters, and Pool 3's servers are configured with the Pool3: Cisco IPMA Server (Primary/Backup) IP Address advanced service parameters.

Figure 18-11 Multiple Active Mode with Unified CM Assistant Server Pools



The Cisco Unified CM Assistant application interacts with the CTIManager for line monitoring and phone control. Each line (including Intercom lines) on a Unified CM Assistant or Manager phone requires a CTI line from the CTIManager. In addition, each Unified CM Assistant route point requires a

CTI line instance from the CTIManager. When you configure Unified CM Assistant, the number of required CTI lines or connections must be considered with regard to the overall cluster limit for CTI lines or connections. (For more information on CTI connection limits per cluster, see Capacity Planning for CTI, page 9-32.) If additional CTI lines are required for other applications, they can limit the capacity of Unified CM Assistant.

Design Considerations for Unified CM Assistant

Unified CM Assistant has the following limitations with regard to overlapping and shared extensions, which you should keep in mind when planning directory number provisioning:

- With Unified CM Assistant in proxy line mode, the proxy line number(s) on the assistant phone should be unique, even across different partitions.
- With Unified CM Assistant in proxy line mode, two Managers cannot have the same Unified CM Assistant controlled line number (DN), even across different partitions.

When enabling Multiple Active Mode and using more than one Unified CM Assistant server pool, ensure that the appropriate server pool (1 to 3) is selected in the Assistant Pool field under the end user Manager Configuration page so that Managers and Assistants are evenly distributed between the Unified CM Assistant server pools. A Manager's associated Assistant will automatically be assigned to the pool where their Manager is configured.

Unified CM Assistant supports a non-secure or secure connection (Transport Layer Security) to the CTI Manager.

Some Cisco endpoints such as the Cisco TelePresence System EX90 might not support Cisco Unified CM Assistant. For details, refer to the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Unified CM Assistant Extension Mobility Considerations

Unified CM Assistant Managers can use Extension Mobility (EM) to log in to their phones in both proxy-line and shared-lined modes. However, the Manager must be configured as a Mobile Manager under the Cisco Unified CM Assistant Manager configuration page of the End-user Directory. When using EM in conjunction with Unified CM Assistant, users should not be able to log in to more than one phone using EM. This behavior can be enabled/disabled via the EM service parameter Multiple Login Behavior. If multiple EM logins by the same user are required within the cluster, Unified CM Assistant Managers who use EM should be instructed not to log in to multiple phones. Allowing a manager to log in to two different phones with EM violates the previously stated restriction that, in proxy line mode, two Managers cannot have the same Unified CM Assistant controlled line number (DN), even across different partitions.

Note Unified CM Assistants cannot use EM to log in to their phones because there is no concept of a Mobile Assistant.

Unified CM Assistant Dial Plan Considerations

Dial plan configuration is extremely important for Unified CM Assistant configured in proxy line mode. To ensure that calls to Manager DNs are intercepted by the Unified CM Assistant RP and redirected to the Assistant phone, calling search spaces and partitions must be configured in such a way that Manager DNs are unreachable from all devices except the Unified CM Assistant RP and the Manager's proxy line on the Assistant phone.

Figure 18-12 shows an example of a proxy line mode Unified CM Assistant dial plan with the minimum requirements for calling search spaces, partitions, and the configuration of various types of devices within these dial plan components. Three partitions are required for proxy line mode, and for the example in Figure 18-12 they are as follows:

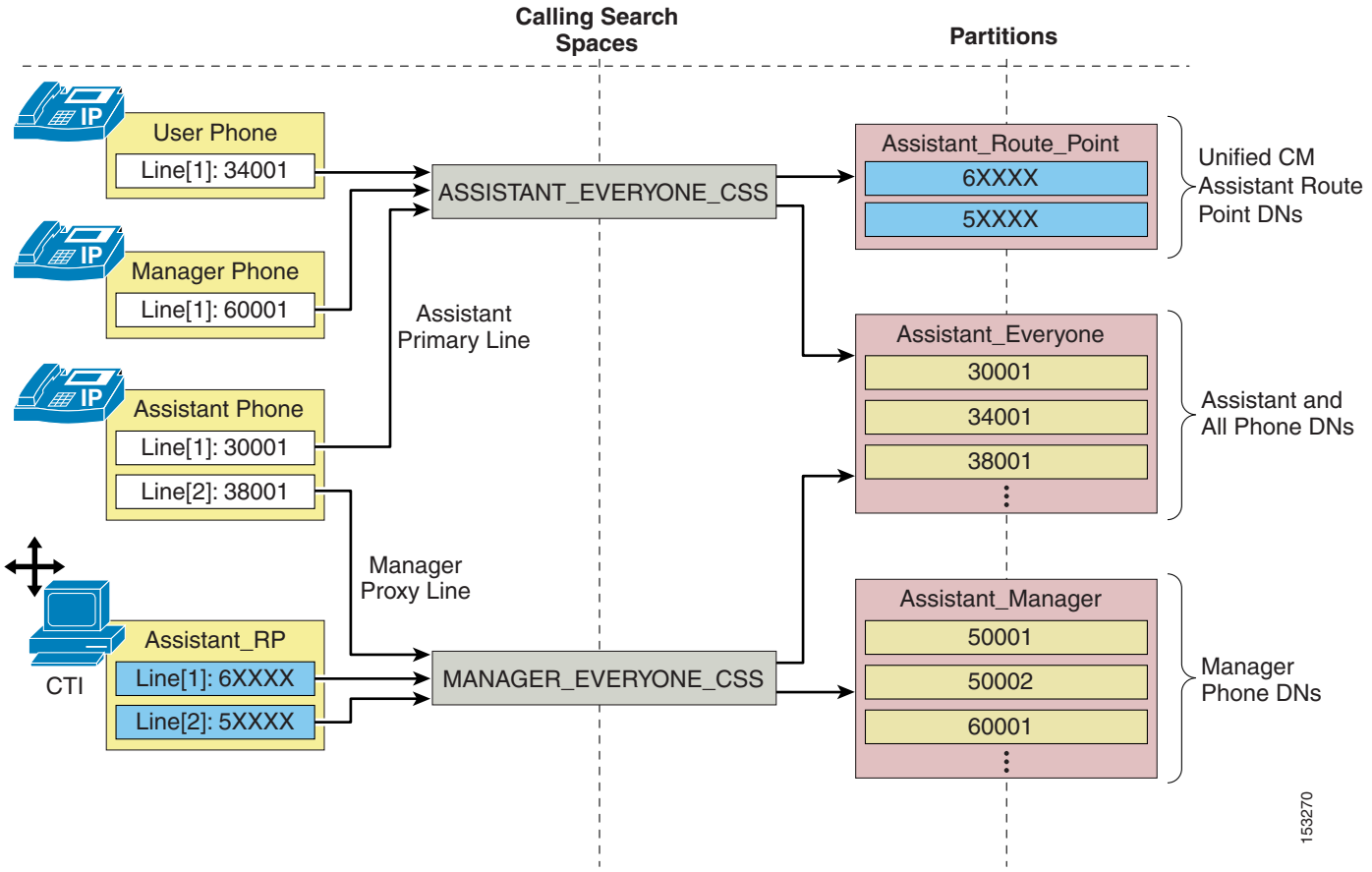
- Assistant_Route_Point partition, containing all the Unified CM Assistant RP DNs
- Assistant_Everyone partition, containing all the Assistant and other user phone DNs
- Assistant_Manager partition, containing all the Manager phone DNs

In addition, two calling search spaces are required, and for the example in Figure 18-12 they are as follows:

- ASSISTANT_EVERYONE_CSS calling search space, containing both the Assistant_Route_Point and Assistant_Everyone partitions.
- MANAGER_EVERYONE_CSS calling search space, containing both the Assistant_Manager and Assistant_Everyone partitions.

That is the extent of the dial plan for this example. However, it is also important to properly configure the various phone and Unified CM Assistant RP DNs or lines with the appropriate calling search spaces so that call routing works as required. In this case all user, Assistant primary (or personal), and Manager phone lines would be configured with the ASSISTANT_EVERYONE_CSS calling search space so that all of these lines can reach all the DNs in the Assistant_Everyone and Assistant_Route_Point partitions. Intercom lines and any other lines configured on devices within the telephony network would be configured with this same calling search space. All Manager proxy lines and all Assistant_RP lines are configured with the MANAGER_EVERYONE_CSS calling search space so that all of these lines can reach the Manager DNs in the Assistant_Manager partition as well as all the DNs belonging to the Assistant_Everyone partition. In this way, the dial plan ensures that only the Assistant_RP lines and the Manager proxy lines on the Assistant phones are capable of reaching the Manager phone DNs directly.

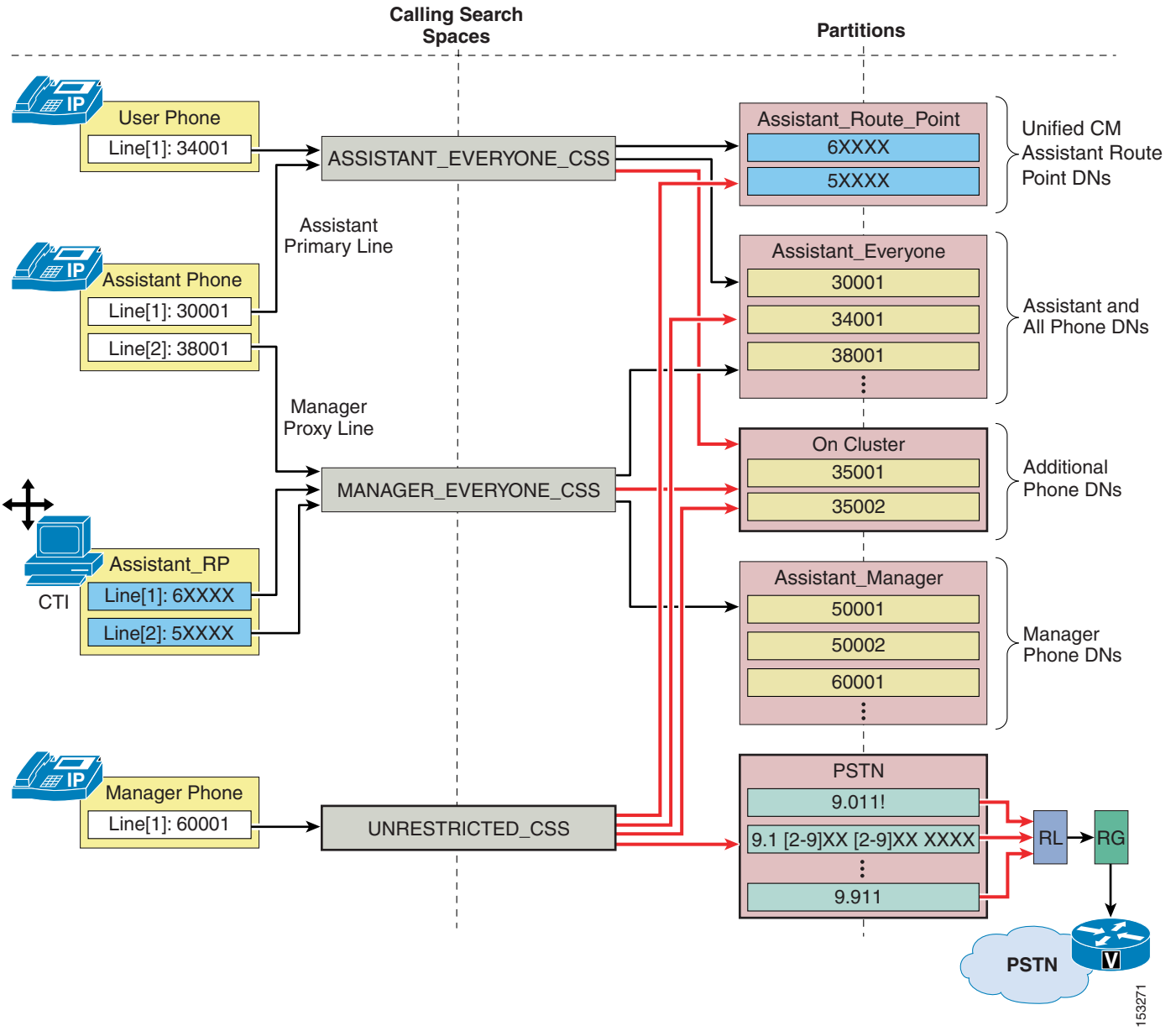
Figure 18-12 Unified CM Assistant Proxy Line Mode Dial Plan Example



The example in Figure 18-12 shows the minimum dial plan requirements for Unified CM Assistant in proxy line mode. However, most real-world telephony networks will have additional or existing dial plan requirements that must be integrated with the Unified CM Assistant calling search spaces and partitions. Figure 18-13 illustrates such an integration dial plan. In this example, the previously discussed dial plan must now handle two additional partitions and an additional calling search space. The On Cluster partition has been added in Figure 18-13, and it contains some additional phone DNs. The On Cluster partition has been added to both of the existing Unified CM Assistant calling search spaces (ASSISTANT_EVERYONE_CSS and MANAGER_EVERYONE_CSS) so that existing devices can reach these added DNs. The UNRESTRICTED_CSS calling search space has also been added to the existing dial plan. This calling search space is configured with the Assistant_Route_Point, Assistant_Everyone, and the recently added On Cluster partitions. In addition, a second new partition called PSTN has been added, and it contains a set of route patterns used for routing calls to the PSTN via the common route list (RL), route group (RG), and voice gateway mechanism. This PSTN partition is configured as part of the UNRESTRICTED_CSS calling search space.

Phone and device line calling search space configurations may be adjusted to incorporate the newly added partitions and calling search spaces, provided the Assistant_RP and Assistant phone Manager proxy lines remain assigned to the MANAGER_EVERYONE_CSS calling search space. In this example, the Manager phone line has been moved from the originally configured ASSISTANT_EVERYONE_CSS calling search space to the new UNRESTRICTED_CSS because it is likely that a Manager would be given unrestricted access to the PSTN.

Figure 18-13 Unified CM Assistant Proxy Line Mode Dial Plan Integration Example



As Figure 18-13 illustrates, integrating additional partitions and calling search spaces into a new or existing Unified CM Assistant dial plan is feasible, but care must be taken to ensure that the underlying proxy line mode mechanism remains intact.

For Unified CM Assistant shared line mode, no special dial plan provisioning is required. Manager and Assistant phones can be configured with calling search spaces and partitions like any other phones in the network because there are no Unified CM Assistant RPs or proxy lines to be concerned about. The only requirement with regard to shared line mode is that the Manager and Assistant DNs must be in the same partition so that shared line functionality is possible.

Unified CM Assistant Console

The Unified CM Assistant Console desktop application or the Unified CM Assistant Console phone service is required in order for assistants to handle calls on a manager's behalf. The desktop application provides assistants with a graphical interface for handling calls, while the phone service provides a menu-driven interface for handling calls. Both the desktop application and the IP phone service allow the assistant to configure the Manager phone and environment and monitor line status and availability. In addition, the desktop application provides other functions such as click-to-call speed dialing and directory entries, which can also be performed on the assistant phone using the traditional softkey and menu approach.

Unified CM Assistant Console Installation

The Unified CM Assistant Console desktop application can be installed from the following URL:

https://<*Server_IP-Address*>:8443/plugins/CiscoUnifiedCallManagerAssistantConsole.exe

(where <*Server_IP-Address*> is the IP address of any node in the cluster)

The Unified CM Assistant Console phone service does not require any installation. To enable the Assistant's phone as a console, subscribe the phone to the Unified CM Assistant phone service. (This is the same service to which Manager phones must also be subscribed.)

Unified CM Assistant Desktop Console QoS

After installation, and in order to handle calls on a Manager's behalf, the Assistant must log on to the application by providing userID and password (as configured in the End-user directory on Unified CM) and will have to toggle status to "online" by clicking the Go Online icon or menu item. Once the user is logged in and online, the desktop application communicates with the Unified CM Assistant server at TCP port 2912. The application chooses an ephemeral TCP port when sourcing traffic. Because the Unified CM Assistant server on Unified CM interfaces with the desktop application for call control (generation and handling of call flows), traffic sourced from Unified CM on TCP port 2912 is QoS-marked by Unified CM as Differentiated Services Code Point (DSCP) of 24 or Per Hop Behavior (PHB) of CS3. In this way, Unified CM Assistant phone control traffic can be queued throughout the network like all other call signaling traffic.

In order to ensure symmetrical marking and queuing, the Unified CM Assistant Console application traffic destined for Unified CM TCP port 2912 should also be marked as DSCP 24 (PHB CS3) to ensure this traffic is placed in the appropriate call signaling queues along the network path toward Unified CM and the Unified CM Assistant server. The Unified CM Assistant Console application marks all traffic as best-effort. This means that you will have to apply an access control list (ACL) at the switch port level (or somewhere along the network path, preferably as close to the console PC as possible) to remark traffic sent by the application PC destined for Unified CM on TCP port 2912 from DSCP 0 (PHB Best Effort) to DSCP 24 (PHB CS3).

Unified CM Assistant Console Directory Window

The directory window within the Assistant Console desktop application enables an assistant to search for end-users in the Unified CM Directory. Search strings entered into the Name field of the directory window are sent to the Unified CM Assistant server, and searches are generated directly against the Unified CM database. Responses to search queries are then sent back to the desktop application by the Unified CM Assistant server.

While the additional traffic generated by directory searches within the desktop application is nominal, this traffic can be problematic in centralized call processing deployments when one or more Unified CM Assistant console applications are running at remote sites. A directory search resulting in a single entry generates approximately one (1) kilobit of traffic from the Unified CM Assistant server to the desktop application. Fortunately, a maximum of 25 entries can be retrieved per search, meaning that a maximum of approximately 25 kilobits of traffic can be generated for each search made by the desktop application. However, if directory searches are made by multiple Unified CM Assistant Console desktop applications across low-speed WAN links from the Unified CM Assistant server, the potential for congestion, delay, and queuing is increased. In addition, directory retrieval traffic is sourced from Unified CM on TCP port 2912, like all other Unified CM Assistant traffic to the desktop. This means that directory retrieval traffic is also marked with DSCP 24 (PHB CS3) and therefore is queued like call signaling traffic. As a result, directory retrieval could potentially congest, overrun, or delay call control traffic.

Note If a directory search generates more than 25 entries, the assistant is warned via a dialog box with the message: “Your search returned more than 25 entries. Please refine your search.”

Given the potential for network congestion, Cisco recommends that administrators encourage Unified CM Assistant Console users to do the following:

- Limit their use of the directory window search function.
- To reduce the number of entries returned, enter as much information as possible in the Name field and avoid wild-card or blank searches when using the feature.

These recommendations are especially important if either of the following conditions is true:

- There are many Unified CM Assistant Assistants within the cluster.
- There are many assistants separated from the Unified CM and/or Unified CM Assistant servers by low-speed WAN links.

Unified CM Assistant Phone Console QoS

In order to handle calls on a Manager's behalf using the Unified CM Assistant Phone Console phone service, the Assistant must log on to the service by providing a userID and PIN (as configured in the End-user directory on Unified CM). Once the user is logged in, the phone console service communicates with Unified CM using HTTPS and SCCP. Call control traffic for Unified CM Assistant call generation and call handling is sent between the phone and Unified CM using SCCP. By default this traffic is marked as Differentiated Services Code Point (DSCP) of 24 or Per Hop Behavior (PHB) of CS3, thus ensuring it is queued throughout the network as call signaling traffic, therefore no additional QoS configuration or marking is required.

WebDialer

WebDialer is a click-to-call application for Unified CM that enables users to place calls easily from their PCs using any supported phone device. There is no requirement for administrators to manage CTI links or build JTAPI or TAPI applications because Cisco WebDialer provides a simplified web application and HTTP or Simple Objects Access Protocol (SOAP) interface for those who want to provide their own user interface and authentication mechanisms.

This section examines the following design aspects of the WebDialer feature:

- WebDialer Architecture, page 18-34
- High Availability for WebDialer, page 18-39
- Capacity Planning for WebDialer, page 18-40
- Design Considerations for WebDialer, page 18-41

WebDialer Architecture

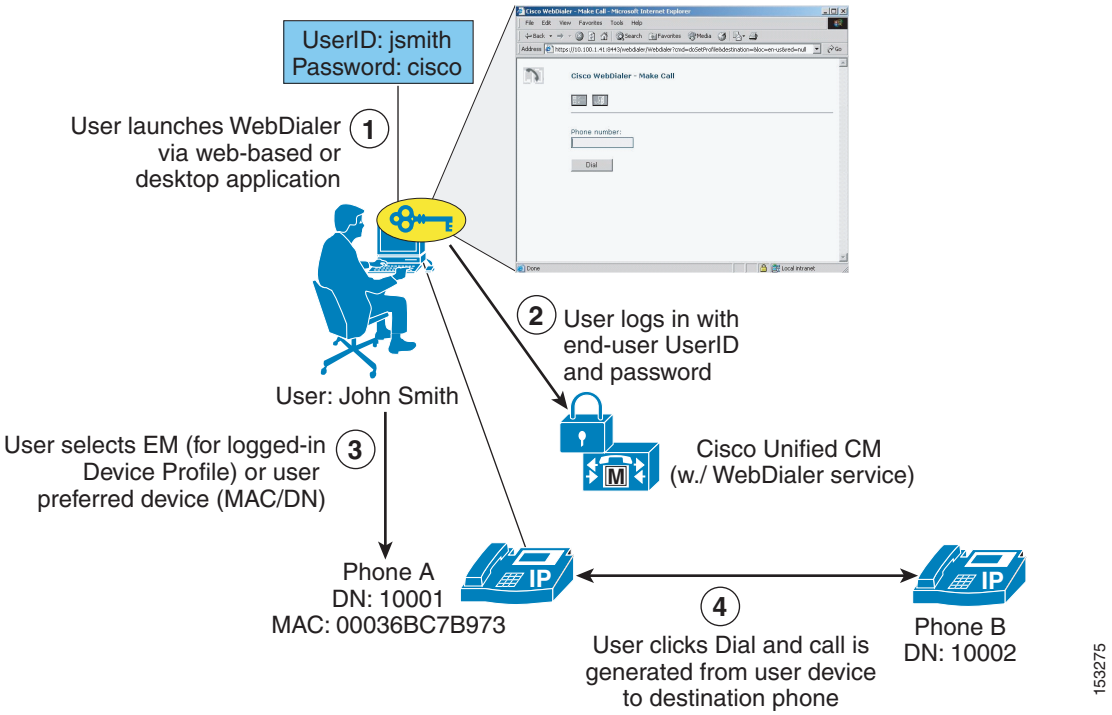
The WebDialer application contains two servlets: the WebDialer servlet and the Redirector servlet. Both servlets are enabled when the Cisco WebDialer Web service is activated on a subscriber server. While related, they each serve different functions and can be configured to run simultaneously.

WebDialer Servlet

Figure 18-14 illustrates a simple WebDialer example. In this example, user John Smith launches WebDialer from a web-based or desktop application (step 1). WebDialer responds with a request for login credentials. The user must respond with a valid userID and password as configured in the Unified CM end-user directory. In this case, John Smith submits userID = jsmith and password = cisco (step 2). Next, based on this login, WebDialer responds with the Cisco WebDialer Preferences configuration page, and the user must indicate either “User preferred device” or “Use Extension Mobility” (assuming the user has an EM device profile). In this case, user John Smith selects “User preferred device” and selects the appropriate MAC address (SEP00036BC7B973) and directory number (10001) for his phone from drop-down menus on the configuration page (step 3). Finally, the user is presented with a screen requesting the phone number to be called (this value may already be indicated) and must click Dial. In this case, John Smith enters 10002 and, after clicking Dial, a call is automatically generated from his phone to Phone B at number 10002 (step 4).

Note If the user has previously logged in to the WebDialer application and a web browser and server cookie are still active, the user will not be prompted to log in again during subsequent requests. The user will be prompted to log in again when the cookie has been cleared at the browser or by a restart of the WebDialer server. Alternatively, the user web browser cookie can be set to expire automatically after a certain number of hours as configured by the User Session Expiry WebDialer service parameter.

Figure 18-14 WebDialer Servlet Operation



Redirector Servlet

The Redirector servlet provides WebDialer functionality in a multi-cluster or distributed call processing environment. This functionality allows the use of a single enterprise-wide web-based WebDialer application between all Unified CM clusters. Figure 18-15 illustrates the basic operation of the Redirector servlet as part of the WebDialer application. In this example, the enterprise has three Unified CM clusters: New York, Chicago, and San Francisco. All three clusters have been configured with a single WebDialer application. The San Francisco cluster has been designated as the Redirector.

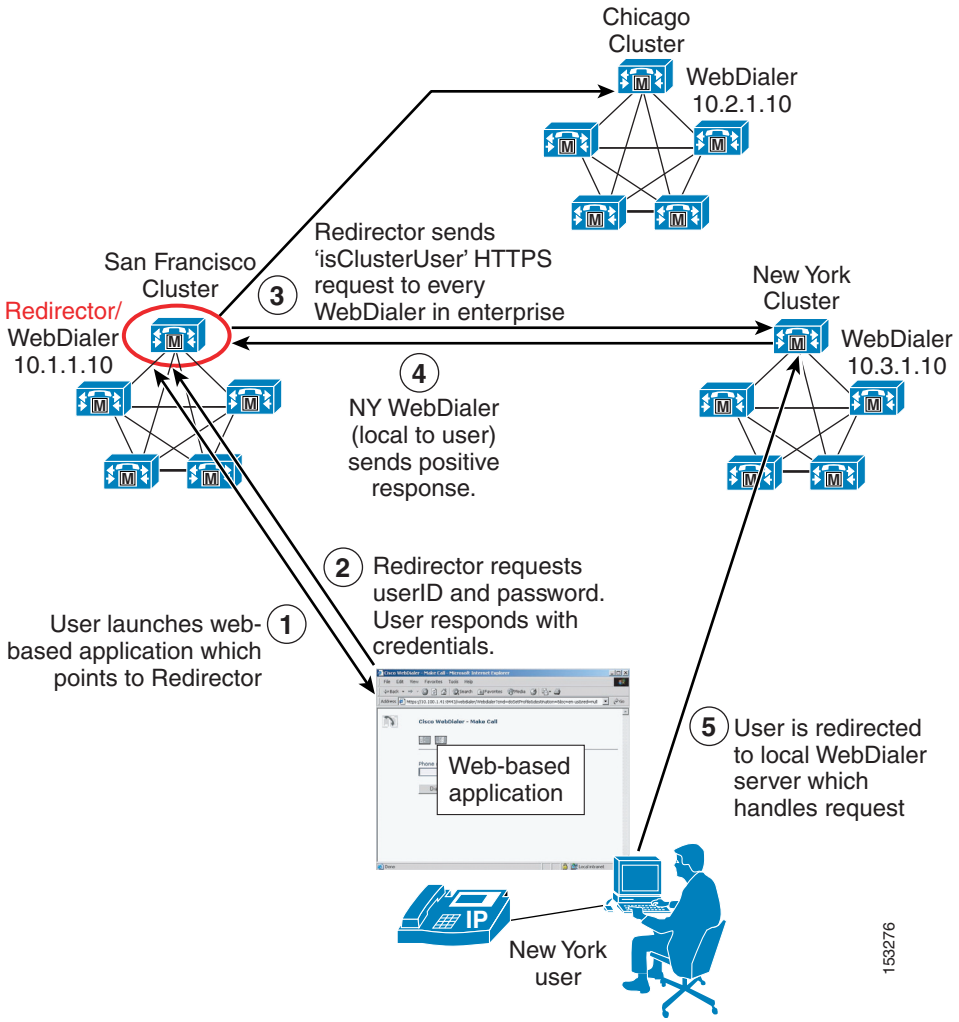
The enterprise-wide web-based application points to the San Francisco Redirector and is launched by the New York user (see step 1 in Figure 18-15). Next the Redirector requests user login, and the New York user responds back with their userID and password (see step 2 in Figure 18-15).

Note If the user has previously logged in to the WebDialer application and a web browser and server cookie are still active, the user will not be prompted to log in again during subsequent requests. Alternatively, the user web browser cookie can be set to expire automatically after a certain number of hours as configured by the User Session Expiry WebDialer service parameter.

The Redirector then broadcasts an isClusterUser HTTPS request to every WebDialer in the enterprise simultaneously (as configured in the List of WebDialers service parameter). In this example, the requests go to the Chicago and New York WebDialer servers (see step 3 in Figure 18-15). Because the New York user is local to the New York cluster, the New York WebDialer responds with a positive response (see step 4 in Figure 18-15). Finally, the New York user is redirected to their local WebDialer server, which will handle the application request (see step 5 in Figure 18-15). The user is not notified of the redirect; however, the URL in the browser address bar will be changed as the user is redirected from the

Redirector to the local WebDialer server). In this example, only one Redirector is deployed; but in order to provide redundancy for the Redirector, configure the Redirector on multiple clusters, as discussed in the section on Service and Component Redundancy, page 18-40.

Figure 18-15 IRedirector Servlet Operation



Note Because the Redirector application is an enterprise-wide application that requires user authentication against the Unified CM Database, Cisco highly recommends that all end-user userIDs be unique across all Unified CM clusters. If they are not, then it is possible that more than one positive response to the isClusterUser request could be received by the Redirector application. If this happens, the user will be asked by the Redirector application to select their local WebDialer server manually. The user will then have to know which server is their local server. If the wrong server is chosen, the WebDialer request will fail.

WebDialer Architecture

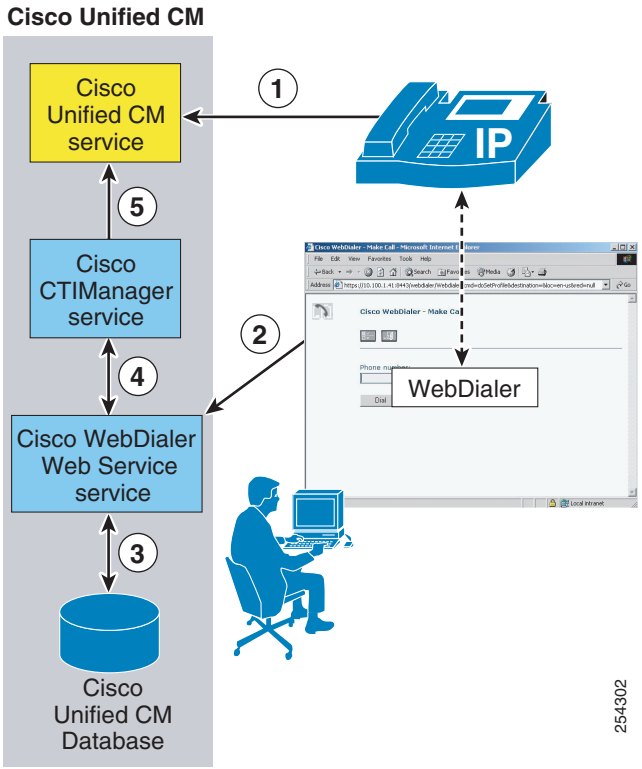
The architecture of the WebDialer application is as important to understand as its functionality. Figure 18-16 depicts the message flows and architecture of WebDialer. The following sequence of interactions and events can occur:

1. WebDialer user phones register and make and receive calls via the Cisco CallManager service (see step 1 in Figure 18-16).
2. The WebDialer application on the user's PC communicates with the Cisco WebDialer Web Service (see step 2 in Figure 18-16) via one of the following interfaces:
   - HTML over HTTPS

     This interface is used by web-based applications based on the HTTPS protocol. This is the only interface that provides access to the WebDialer and Redirector servlets.
   - Simple Object Access Protocol (SOAP) over HTTPS

     This interface is used by desktop applications based on the SOAP interface.
3. The WebDialer Web service reads user and phone information from the Unified CM Database (see step 3 in Figure 18-16).
4. The WebDialer Web service in turn interacts with the CTIManager service for exchanging line and phone control information (see step 4 in Figure 18-16).
5. The CTIManager service passes WebDialer phone control information to the Cisco CallManager service (see step 5 in Figure 18-16).

Figure 18-16 WebDialer Architecture



Note Although Figure 18-16 shows the Cisco CallManager, CTIManager, and WebDialer Web Service services all running on the same node, this configuration is not a requirement. These services can be distributed among multiple nodes in the cluster, but they are shown on the same node here for ease of explanation.

WebDialer URLs

The WebDialer application can be accessed from web-based applications via the HTML-over-HTTPS interface using the following URLs:

- WebDialer servlet

  https://<*Server-IP_Addr*>:8443/webdialer/Webdialer?destination=<*Number_to_dial*>

  (where <*Server_IP-Address*> is the IP address of any node in the cluster running the Cisco WebDialer Web Service service, and where <*Number_to_dial*> is the number that the WebDialer user wishes to dial)

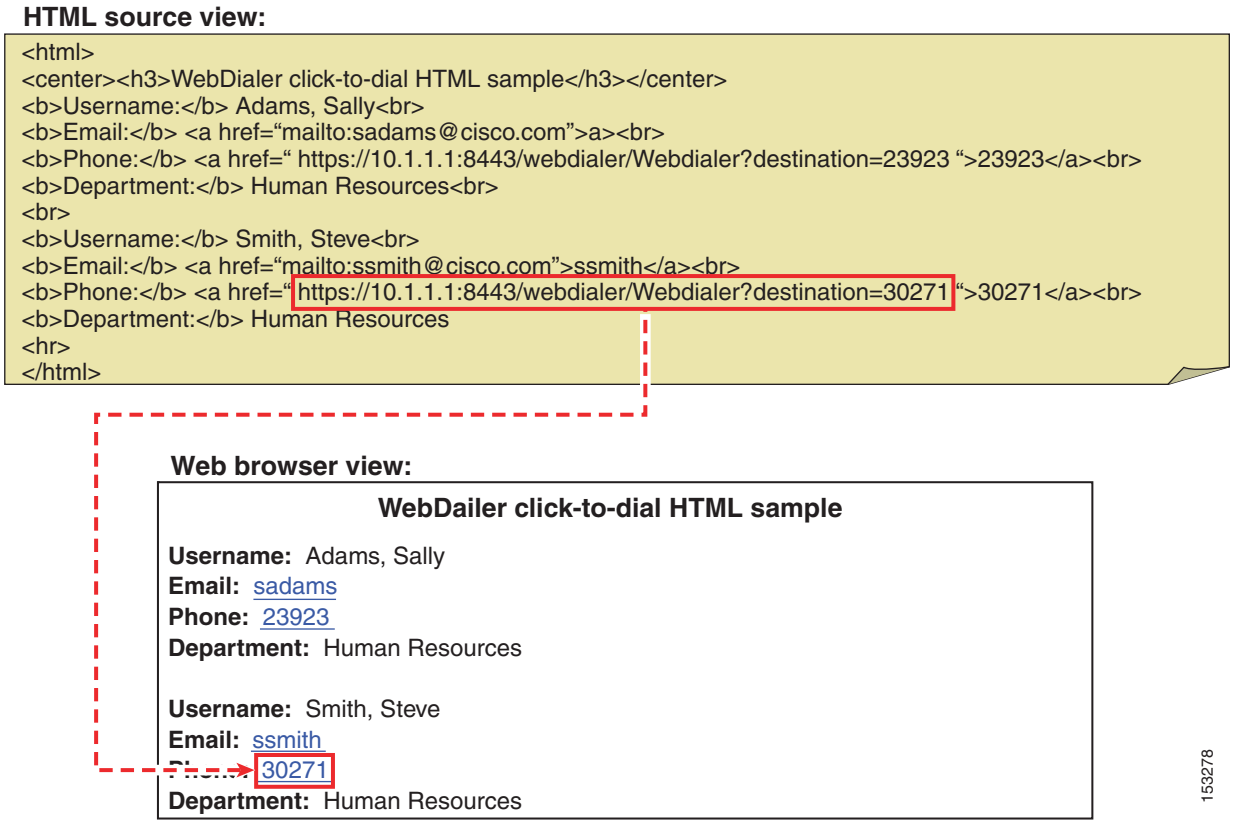
- Redirector servlet

  https://<*Server-IP_Addr*>:8443/webdialer/Redirector?destination=<*Number_to_dial*>

  (where <*Server_IP-Address*> is the IP address of any node in the enterprise running the Cisco WebDialer Web Service service, and where <*Number_to_dial*> is the number that the WebDialer user wishes to dial)

Figure 18-17 gives an example of HTML source code used in a click-to-call web-based application calling the Cisco WebDialer application. In this example, the URL https://10.1.1.1:8443/webdialer/Webdialer?destination=30271 in the HTML source view corresponds to the "Phone: 30721" link for user Steve Smith within the web browser view. A user clicking on this link would launch the WebDialer application and, after logging in and clicking Dial, would generate a call from the user's phone to Steve Smith's phone. The same code could be used for a click-to-call application using the Redirector function by changing the URL to https://10.1.1.1:8443/webdialer/Redirector?destination=30271.

Figure 18-17 WebDialer URL HTML Example



For information and examples of SOAP-over-HTTPS source code to be used in click-to-call desktop applications, refer to the WebDialer API Programming information in the *Cisco WebDialer Developer Guide*, available at

https://developer.cisco.com/site/webdialer/discover/getting-started/

High Availability for WebDialer

WebDialer application redundancy can be provided at two levels:

- Redundancy at the component and service level

  At this level, redundancy must be considered with regard to WebDialer and CTIManager service redundancy. Likewise, the lack of publisher redundancy and the impact of this component failing should also be considered.

- Redundancy at the device and reachability level

  At this level, redundancy should be considered as it relates to user phones and the WebDialer user interface.

Service and Component Redundancy

As shown in Figure 18-16, WebDialer functionality is primarily dependent on the Cisco WebDialer Web Service and the Cisco CTIManager services. The WebDialer service can be enabled on multiple nodes within the cluster. Reachability to those multiple nodes is described in the section on Device and Reachability Redundancy, page 18-40. In the case of CTIManager, redundancy is automatically built-in using a primary and backup mechanism. Two CTIManager servers or services can be defined within the cluster using the Primary Cisco CTIManager and the Backup Cisco CTIManager service parameters. By configuring these parameters, you can make the CTIManager service redundant. Thus, if the primary CTIManager fails, CTIManager services can still be provided by the backup CTIManager. If the WebDialer server to which the web-based (or desktop) application is pointing fails and the primary and backup CTIManager services on cluster nodes also fail, the WebDialer application will fail. The WebDialer service is not dependant upon the Unified CM publisher

Device and Reachability Redundancy

Redundancy for WebDialer at the device level relies on a number of mechanisms. First and foremost, user phones rely on the built-in redundancy provided by a combination of the device pool and Unified CM group configuration for device registration.

The WebDialer service can run on multiple Unified CM subscribers in the same cluster to provide redundancy, however many applications might not be equipped to handle more than one IP address. Cisco recommends using a Server Load Balancer (SLB) to mask the presence of multiple WebDialer servers in the enterprise. SLB functionality provides a virtual IP address or DNS-resolvable hostname that front-ends the real IP addresses of the WebDialer servers. Most SLB devices, such as a Cisco device running the Cisco IOS SLB feature, can be configured to monitor the status of multiple WebDialer servers and automatically redirect requests during failure events. The SLB feature can also be configured to load-balance WebDialer requests when additional click-to-call capacity is required. As an alternative, DNS Service (SRV) records can also be used to provide redundancy.

Similarly in a multicluster environment, if a single Redirector servlet is supporting multiple WebDialers, it could be a single point of failure. To avoid this single point of failure, configure Redirector servlets for each cluster and use a Server Load Balancer (SLB) to provide a virtual IP address or DNS-resolvable hostname that front-ends the real IP addresses of the Redirector servers.

Capacity Planning for WebDialer

The WebDialer and Redirector services can run on one or more subscriber nodes within a Unified CM cluster, and they support the following capacities:

- Each WebDialer service can handle up to 4 call requests per second per node.
- Each Redirector service can handle up to 8 call requests per second.

The following general formula can be used to determine the number of WebDialer calls per second (cps):

(Number of WebDialer users) * ((Average BHCA) / (3600 seconds/hour))

When performing this calculation, it is important to estimate properly the number of BHCA per user that will be initiated specifically from using the WebDialer service. The following example illustrates the use of these WebDialer design calculations for a sample organization.

Example 18-1 Calculating WebDialer Calls per Second

Company XYZ wishes to enable click-to-call applications using the WebDialer service, and their preliminary traffic analysis resulted in the following information:

- 10,000 users will be enabled for click-to-call functionality.
- Each user averages 6 BHCA.
- 50% of all calls are dialed outbound, and 50% are received inbound.
- Projections estimate 30% of all outbound calls will be initiated using the WebDialer service.

Note These values are just examples used to illustrate a WebDialer deployment sizing exercise. User dialing characteristics vary widely from organization to organization.

10,000 users each with 6 BHCA equates to a total of 60,000 BHCA. However, WebDialer deployment sizing calculations must account for placed calls only. Given the initial information for this sizing example, we know that 50% of the total BHCA are placed or outbound calls. This results in a total of 30,000 placed BHCA for all the users enabled for click-to-call using WebDialer.

Of these placed calls, the percentage that will be initiated using the WebDialer service will vary from organization to organization. For the organization in this example, several click-to-call applications are made available to the users, and it is projected that 30% of all placed calls will be initiated using WebDialer.

(30,000 placed BHCA) * 0.30 = 9,000 placed BHCA using WebDialer

To determine the number of WebDialer servers required to support a load of 9,000 BHCA, we convert this value to the average call attempts per second required to sustain this busy hour:

(9,000 call attempts / hour) * (hour/3600 seconds) = 2.5 cps

Each WebDialer service can support up to 4 cps, therefore one node can be configured to run the WebDialer service in this example. This would allow for future growth of WebDialer usage. In order to maintain WebDialer capacity during a server failure, additional backup WebDialer servers should be deployed to provide redundancy.

Keep in mind that the Cisco WebDialer application interacts with the CTIManager for phone control. When enabled, each WebDialer service opens a single persistent CTI connection to the CTIManager. In addition, each WebDialer individual MakeCall (or EndCall) request generates a temporary CTI connection. The number of CTI connections required to handle WebDialer call rates also applies against the CTI connection limits per cluster. (For more information on CTI connection limits per cluster, see Capacity Planning for CTI, page 9-32.)

Design Considerations for WebDialer

The following guidelines and restrictions apply with regard to deployment and operation of WebDialer within the Unified CM environment:

- The administrator should ensure that all WebDialer users are associated with a phone or device profile in the Unified CM end-user directory.
  - If the user selects "Use permanent device" under the Cisco WebDialer Preferences screen with no phone association, then the following message is received when the Dial button is pressed:

    “No supported device configured for user”

  - If the user selects Use Extension Mobility under the Cisco WebDialer Preferences screen with no device profile association (or the user is not logged in using a profile), then the following message is received when the Dial button is pressed:

    "Call to <*dialed_ number*> failed: User not logged in on any device"

- An application interfaces with the WebDialer and Redirector servlets through HTTPS.
- If using Client Matter Codes (CMC) or Forced Authorization Codes (FAC), WebDialer users must enter the proper code at the tone by using the phone's keypad. Failure to enter the appropriate code at the tone will result in call failure signaled by a reorder tone.
- Cisco WebDialer is available on any Cisco endpoints that support Cisco Computer Telephony Integration (CTI).

For a list of Cisco endpoints that support Cisco Computer Telephony Integration (CTI), refer to the *CTI (TAPI/JTAPI) Supported Device Matrix*, available at

https://developer.cisco.com/site/jtapi/wiki/cti-tapi-jtapi-supported-device-matrix/

Cisco Unified Attendant Consoles

Attendant console integrations enable a receptionist to answer and transfer or dispatch calls within an organization from a Microsoft Windows desktop application designed specifically for this purpose. Cisco Unified Attendant Consoles provide access to local and corporate directories, where uses can view user line status, Jabber status, and Skype for Business status (Cisco Unified Attendant Console Advanced only).

The Cisco Unified Communications portfolio provides two editions of Cisco Unified Attendant Console:

- Cisco Unified Attendant Console Standard

  This console is a local installation that integrates directly with Unified CM to monitor and control users' phones (the "serverless" solution). Call routing and controls executed via the console application mimic that of the device used to log in to the software.

- Cisco Unified Attendant Console Advanced

  This local attendant console application connects to a central Cisco Unified Attendant Console Advanced Windows server application (separate from Unified CM). The Cisco Unified Attendant Console Advanced server communicates with Unified CM through CTI and AXL over Secure Socket Layer (SSL). All call routing and controls are executed by the central server application.

This section examines the following design aspects of the attendant consoles:

- Design Considerations for Cisco Unified Attendant Console Standard, page 18-43
- Cisco Unified Attendant Console Advanced Architecture, page 18-43
- Design Considerations for Cisco Unified Attendant Console Advanced, page 18-45
- Capacity Planning for Cisco Unified Attendant Consoles, page 18-47

Design Considerations for Cisco Unified Attendant Console Standard

The following design guidelines and restrictions apply to the deployment and operation of Cisco Unified Attendant Console Standard within the Unified CM telephony environment.

- Each installation of Cisco Unified Attendant Console Standard requires a CTI and AXL connection with Unified CM.
- Console user device(s) and all devices belonging to contacts – whose line state (busy lamp field) is required within the Attendant Console directory – must be assigned to the defined Unified CM application user.
- The total of lines (the total number of devices and their respective lines) must not exceed 5,000. This is not a hard-coded limit; however, exceeding the limit can degrade performance and stability of the Attendant Console and Unified CM environment.

For design considerations regarding shared lines and call routing, refer to the latest version of the *Cisco Unified Attendant Console Standard Installation and Configuration Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-attendant-console-standard/model.html

Cisco Unified Attendant Console Advanced Architecture

Figure 18-18 illustrates the high-level architecture of a server-based Cisco Unified Attendant Console Advanced integration. Understanding the functionality and operation of the solution enhances the understanding of the architecture itself. The following steps (denoted in Figure 18-18) detail the events involved for a typical call into an attendant console.

1. A call comes into Unified CM, and the called number matches the directory number configured on a CTI route point.
2. The CTI route point is CTI-controlled by the attendant console server application and is associated with a Queue Direct Dial In (DDI) configured on the server.
3. The attendant console server application immediately redirects the call internally to one of its Computer Telephony (CT) Gateway Devices. As part of this process, the attendant console server application sends a CTI redirect message to the CTI Manager service to redirect the call to a CTI port.

Note A CTI redirect message does not result in a connected call; the call is not answered and there is no media connection.

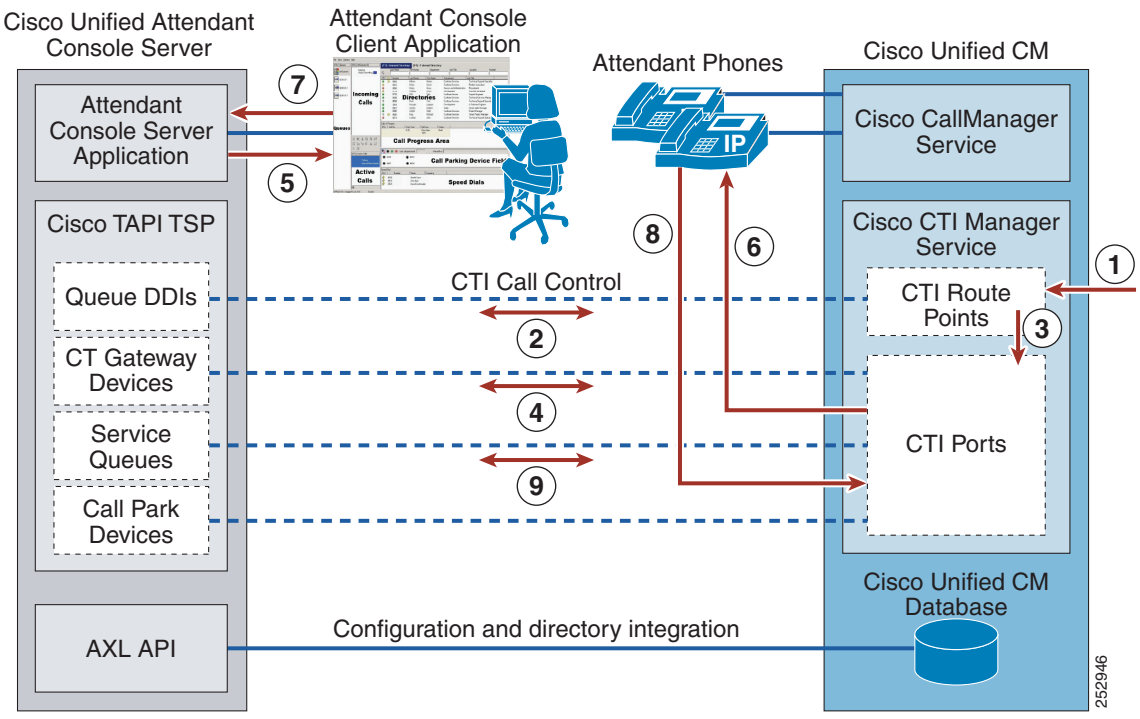
4. The attendant console server application now associates the call with the CT Gateway Device and controls the call on the CTI port.
5. At this point, the call is presented to the attendant console client applications in the system that are associated with the Queue DDI.
6. Once an attendant chooses to answer the call through the attendant console client application, another CTI redirect message is sent to the CTI Manager service, which moves the call from the CTI port to the answering attendant's physical phone. The call is automatically connected on the attendant's phone, either to the handset or the headset, depending on the phone configuration. The region and location settings of the attendant's phone and the initiating gateway or phone dictate the codec used for media.

7. When a transfer to another extension is required, the attendant initiates the transfer through the attendant console client application, which communicates the transfer to the attendant console server application.
8. The attendant console server application internally associates the call with a Service Queue and sends a CTI redirect message to the CTI Manager service. This redirects the call from the attendant's phone to a CTI port controlled by the attendant console server application.

Note A call transfer may also be initiated from the attendant's phone; however, this would remove the attendant console server application from the call flow, and enhanced functionality (such as the transfer recall feature) would no longer be possible.

9. At this stage, the Service Queue actually answers the call (there is a short connect) before issuing the transfer, therefore the Cisco Media driver installed on the attendant console server application is invoked. The region and location settings of this CTI port and the call-initiating gateway or phone dictate the codec used for media. The configured Music on Hold (MoH) audio sources of the CTI port also affect the MoH heard by the caller. Transfers are performed in this manner so that the attendant console client application still maintains control of the call if there is no answer. Once the call is received by the final party, the attendant console server application is removed from the call flow.

Figure 18-18 Architecture for Server-Based Cisco Unified Attendant Consoles



The attendant console server application's call park function does not use the inherent call park feature of Unified CM. Instead, it uses its own call park facility using Call Park Devices. Call Park Devices work very much like the Service Queues as outlined in steps 7 to 9 of Figure 18-18. Similar to transfers, Call Park Devices allow the attendant console server application to maintain control of the call for the duration of the parked call.

High Availability for Cisco Unified Attendant Console Advanced

Cisco Unified Attendant Console Advanced can be installed in a resilient configuration with two Cisco Unified Attendant Console servers:

- Publisher — The primary server used by the clients. If this server fails, all attendant operators are switched to the subscriber server. Once the publisher is running again, the operators are prompted to reconnect (or are automatically reconnected, if so configured) to the publisher.
- Subscriber — Used if the publisher stops running for any reason.

You should consider providing redundancy on both sides of the integration for both CTI and AXL communication.

Regarding CTI, the attendant console server application uses the Cisco TAPI Telephony Service Provider (TSP) plug-in (downloaded from Unified CM) to communicate with the CTI Manager service. Cisco TSP allows for the configuration of a primary and backup CTI Manager service. Cisco recommends enabling the CTI Manager service on at least two Unified CM subscriber nodes in the cluster to gain resilience in case the primary CTI Manager service goes offline. In the event of an attendant console server failure, resilience is achieved by leveraging the Call Forward Unregistered (CFU) and Call Forward CTI failure destinations configured against the CTI route points associated with Queue DDIs. If the attendant console server application is offline, calls will automatically follow the Call Forward setting. For example, when redundant attendant consoles are deployed, calls will be forwarded to the Cisco Unified Attendant Console subscriber server when the publisher is off-line. With a single attendant console server, the destination could be a Hunt Pilot number or a Directory Number (DN) associated with a single IP phone.

AXL communication is enabled by activating the Cisco AXL Web Service on a Unified CM node. Multiple Unified CM nodes can have the Cisco AXL Web Service enabled, but the attendant console server application has only a single entry for Unified CM connectivity. In the event of a failure, an administrator could update this entry to a backup Unified CM node running the Cisco AXL Web Service. When redundant Cisco Unified Attendant Consoles are deployed, the attendant console servers can be configured on different Unified CM nodes for the AXL Web Service.

The Unified CM has a series of CTI route points and CTI ports belonging to Cisco Unified Attendant Console Advanced. The associated device pool dictates a Unified CM group that contains a prioritized list of the Unified CM call processing nodes responsible for maintaining registration. When the primary Unified CM in the Unified CM group is offline, the CTI route points and CTI ports have the ability to register with a secondary Unified CM node, thus allowing for high availability of the CTI route points and ports themselves.

Design Considerations for Cisco Unified Attendant Console Advanced

The following design guidelines and restrictions apply to the deployment and operation of Cisco Unified Attendant Console Advanced within the Unified CM telephony environment.

- The following general design guidance applies to the attendant console server application components:
  - Queue DDI

    One unique Queue DDI is required for each unique incoming directory number in the system that should be routed specifically to the attendant consoles.
  - CT Gateway Device

    Every incoming call into a Queue DDI is immediately redirected to a CT Gateway Device. Design the system so that the number of CT Gateway Devices can handle the maximum expected number of incoming calls at any given time.

  - Service Queue

    Each time an attendant transfers a call or places a call on hold, a Service Queue is required. The system should be designed so that there are enough Service Queues to sustain the maximum number of calls that all attendants in the system are in the process of transferring or putting on hold at any given time. A general guideline is to provide 3 or 4 Service Queues per attendant, but some scenarios might require more.

  - Call Park Device

    Each time an attendant invokes the Call Park feature through the attendant console client application, a Call Park Device is required. This feature does not use the inherent Call Park capability of Unified CM. Design the system so that there are sufficient Call Park Devices to handle the maximum number of calls parked by all attendants in the system at any given time.

- Every Queue DDI, CT Gateway Device, Service Queue, and Call Park Device configured in the attendant console server application creates a CTI route point or CTI port in Unified CM. The number of CTI connections required to handle the Cisco Unified Attendant Console Advanced integration also counts toward the CTI connection limits per cluster. (For more information on CTI connection limits per cluster, see Capacity Planning for CTI, page 9-32.)
- The attendant console server application provides busy lamp field (BLF) monitoring of end-user devices, but it is important to note that this does not use the same facility in Unified CM that provides BLF speed dial capability. Instead, the attendant console server application communicates through CTI with Unified CM to obtain line state information on monitored devices. Once the attendant console server application monitors an end-user device, it continues monitoring this device through CTI until the number of devices monitored for BLF reaches 2,000 devices. Once this limit is reached, the BLF plug-in begins to drop devices from the list of monitored devices to accommodate the newly requested devices, thus ensuring that the number of devices monitored by the attendant console server through CTI does not exceed the limit (2,000). These devices monitored through CTI also count toward the CTI limits in Unified CM.
- With respect to Quality of Service (QoS), the attendant console server application, the attendant console client application, and the Cisco TSP all send their traffic marked as Best Effort (DSCP=0). If this traffic traverses a WAN or a link that is typically congested, packets must be marked to receive preferential treatment through the network. For a complete list of the TCP port numbers associated with these applications, refer to the latest version of the *Cisco Unified Attendant Console Advanced Administration and Installation Guide*, available at

  https://www.cisco.com/c/en/us/support/unified-communications/unified-attendant-consoles/products-maintenance-guides-list.html
- Cisco TSP is not aware of partitions; therefore, Cisco Unified Attendant Console Advanced does not support overlapping dial plans (this includes Attendant Console system numbers, console user numbers, or contact numbers).

For additional design guidance on Cisco Unified Attendant Console Advanced, refer to the documentation available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-attendant-consoles/products-implementation-design-guides-list.html

Capacity Planning for Cisco Unified Attendant Consoles

For a comparison of the various Cisco Unified Attendant Console editions and their respective capacities, refer to the latest Cisco Unified Attendant Console product documentation available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-attendant-consoles/tsd-products-support-series-home.html

To size a Unified CM cluster properly, your Cisco Partner or Cisco Systems Engineer should use the Cisco Collaboration Sizing Tool (https://www.cisco.com/go/cst) to validate all designs that incorporate a large number of CTI resources and high call volumes, because there are many interdependent variables that can affect Unified CM cluster scalability. The Sizing Tool can accurately determine the number of servers or clusters required to meet your Attendant Console design criteria.

Cisco Paging Server

The Cisco Paging Server allows users to send audio-only messages to groups of up to 50 IP phones in an organization. The Cisco Paging Server is Singlewire InformaCast in basic paging mode. Those who wish to page larger groups of phones or other endpoints or to schedule broadcasts should consider upgrading to advanced notification mode.

The Cisco Paging Server is distributed as an open virtual appliance (OVA) and it runs as a virtual machine within VMware. This virtual machine may run co-resident with the Unified CM virtual machine. The Cisco Paging Server communicates with Unified CM using SIP, SNMP, AXL and CTI. A single Cisco Paging Server per Unified CM cluster is supported.

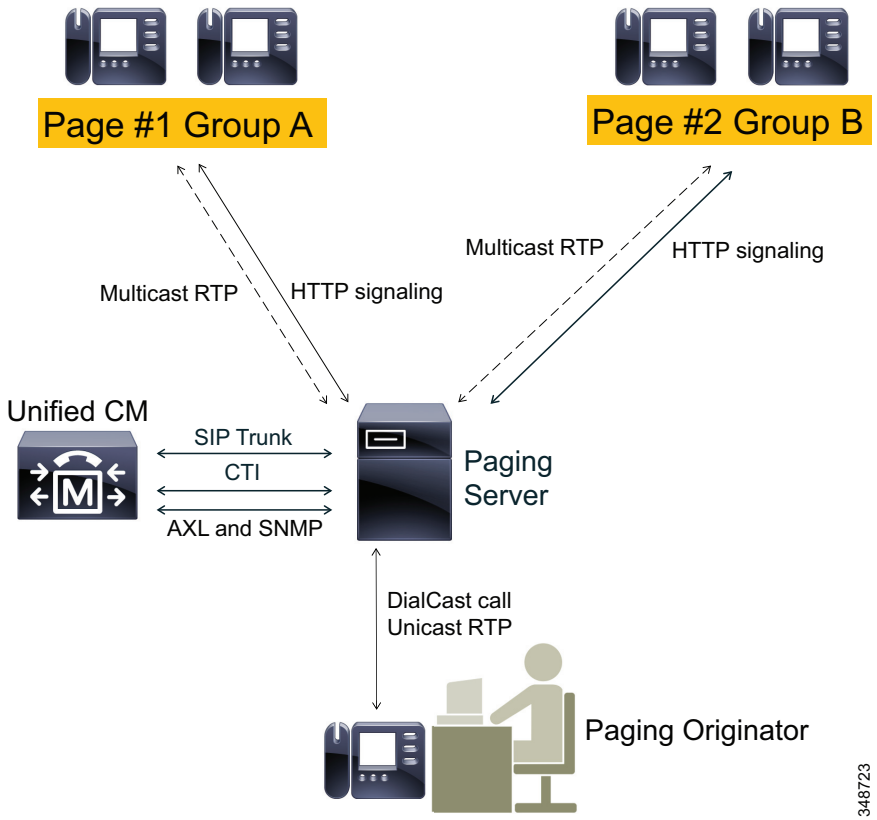
The Cisco Paging Server communicates with IP phones using HTTP. Beginning with Cisco Paging Server 9.0.1, either HTTP or CTI may be used to communicate with phones. In HTTP mode, Cisco Paging Server sends commands and credentials to each IP phone HTTP server. IP phones validate these credentials and then execute the commands. CTI mode sends commands to each phone via Unified CM. In CTI mode, Cisco Paging Server does not need to send credentials with each request, so each phone does not have to activate its web server, and commands are executed more quickly. In addition, CTI mode allows faster checking of busy phones.

When the Cisco Paging Server starts, and at configurable intervals after that, it connects with Unified CM using SNMP. The Cisco Paging Server uses SNMP to find the other Unified CM cluster member IP addresses as well as a list of phones registered to each cluster member. Once the SNMP communications are complete, the Cisco Paging Server uses AXL to determine additional information regarding each registered phone, such as device name, description, device pool, calling search space, directory number, and location. This information can be used to build logical groups of phones, called *recipient groups*. In the Cisco Paging Server, recipient groups can contain a maximum of 50 phones.

Broadcasts are always initiated as voice calls to the Cisco Paging Server. The service that answers these calls on the Cisco Paging Server is called DialCast. DialCast can receive calls through CTI and/or SIP. In the case of CTI, calls arrive and are serviced on a CTI route point (the Cisco Paging Server does not require CTI ports to answer inbound calls). In the case of SIP, calls depart Unified CM on a SIP trunk. Both CTI and SIP are valid and supported. However, Cisco recommends SIP call flows over CTI because troubleshooting SIP is much easier than troubleshooting CTI.

The Cisco Paging Server sends audio to IP phones using multicast. The multicast stream is originated by the Cisco Paging Server and received by the IP phones (see Figure 18-19).

Figure 18-19 Cisco Paging Server Sending Messages to Multiple Groups of Phones



This sequence describes how the Cisco Paging Server initiates a broadcast to one or more IP phones as illustrated in Figure 18-19:

1. The caller dials a predefined number in Unified CM. This number routes the call to the Cisco Paging Server over either a SIP trunk or CTI route point.
2. The Cisco Paging Server answers the call as a DialCast call.
3. The caller hears a low stall tone. While the Cisco Paging Server plays this tone, it is sending a command via HTTP to each phone in the recipient group. The command requests each phone to join the multicast group.
4. Once all phones have joined the multicast group, the Cisco Paging Server plays a high go-ahead tone. When the caller hears this tone, it indicates that the Cisco Paging Server is transmitting the RTP stream from the calling IP phone as a multicast RTP stream out to the receiving phones. When the caller speaks, their voice is sent to the receiving phones.
5. When the caller hangs up, the Cisco Paging Server sends another request to each IP phone, this time to leave the multicast group, and the broadcast is over.

Design Considerations for Cisco Paging Server

Phone Selection and Unified CM Feature Interactions

- Not all Cisco IP phones are compatible with the Cisco Paging Server. For a current list, refer to the compatibility information available at

  https://www.singlewire.com/compatibility-matrix

- To receive broadcasts, IP phones must have the speakerphone enabled.
- The Cisco Paging Server does not recognize Do Not Disturb (DND) and will send to phones with DND enabled.

Multicast Considerations

- If the Cisco Paging Server and IP phones are on separate IP subnets, the routers in between those two subnets must be configured for multicast routing.
- The Cisco Paging Server does not require any particular method of multicast routing (SM, DM, S-DM, SSM, and so forth).
- Some wide area network environments do not support multicast routing. For those environments, GRE tunnels may be built between sites and used to transport multicast.
- The multicast media streams always use the G.711 mu-law codec. No other codecs are allowed or supported.
- The RTP flow is always unicast from the initiator to the Cisco Paging Server, and then multicast from the Cisco Paging Server to the receiving phones. If the Cisco Paging Server is deployed centrally, then pages may cross WAN boundaries.
- Cisco Paging Server multicast media streams are not calls. They do not count against either RSVP or locations-based call admission control. WAN engineers should budget for these multicast media streams in addition to other voice flowing across the enterprise network. However, the call between the initiator and the Cisco Paging Server is a normal voice call. This call is subject to normal call admission control restrictions.
- If you have a multisite deployment, Cisco recommends configuring the Paging Server to use a range of multicast addresses rather than the single default address. The reason for this is that Internet Group Management Protocol (IGMP) multicast joins are effective for the multicast address only, not the address and the port. If two broadcasts are going on simultaneously to two different sites, a phone at either site will send out an IGMP join to the multicast address only. If both sites use the same single address, the RTP streams for both broadcasts will be sent to both sites.
- Multicast streams cannot be encrypted, except when crossing IPSec tunnels.
- Different phone models and firmware versions may use different IGMP versions, which can impact switch configuration.

Other Considerations

- Inbound calls to DialCast must be G.711 mu-law calls. Calls arriving to DialCast using other codecs must be transcoded.
- The Cisco Paging Server does maintain a CTI connection to Unified CM, but the load that this CTI connection places on Unified CM is very low. The resources that this connection requires remain constant regardless of cluster size.
- The Cisco Paging Server requires that any firewalls between the server and the phones not be configured to use Network Address Translation (NAT).
- Beginning in Cisco Paging Server 8.4, QoS values are set to Unified CM default values (DSCP CS3 for signaling and DSCP EF for media). Signaling (DSCP CS3) applies to CTI and SIP traffic, while media (DSCP EF) applies to SIP and CTI-initiated RTP streams as well as outbound multicast RTP streams. Paging Server DSCP values cannot be changed in the field. Customers that wish to use DSCP values different from these must re-mark Paging Server traffic in the network.

For more information, refer to the latest version of the *InformaCast Virtual Appliance Basic Paging Installation and User Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/paging-server/products-maintenance-guides-list.html

Cisco Voice Messaging

Revised: March 1, 2018

This chapter describes the voice messaging solutions available in the Cisco Unified Communications System. It includes the Cisco voice messaging products Cisco Unity Connection and Cisco Unity Express, and it covers the design guidelines and best practices for deploying these products together with Cisco Unified Communications Manager (Unified CM). This chapter also covers aspects of integration with third-party voicemail systems using industry standard protocols.

Although this guide focuses on the messaging deployment scenarios with regard to Unified CM, Cisco Unified Communications Manager Express (Unified CME) is also noted where applicable, especially when used with Survivable Remote Site Telephony (SRST) fallback support in a centralized Unified CM deployment.

This chapter covers the following topics:

- Voice Messaging Portfolio, page 19-2
- Messaging Deployment Models, page 19-3
- Messaging and Unified CM Deployment Model Combinations, page 19-5
- Voicemail Networking, page 19-28
- Best Practices for Voice Messaging, page 19-32
- Third-Party Voicemail Design, page 19-47

The chapter begins with a short description of each of the products in the Cisco messaging solutions portfolio and provides a simple overview of where each product fits in an enterprise Unified Communications solution. Next, messaging deployment models form the basis of discussion for voicemail integrations, which start with a definition of the various messaging deployment models and then explain how each of the messaging deployment models fits into the various Unified CM call processing deployment models. Cisco Unity Connection is discussed in this section, while Cisco Unity Express has a dedicated section for its supported deployment models. Key design guidelines are covered for interoperability available within the Cisco Voice Messaging product portfolio. Virtualization is also covered along with the important design factors to be considered while designing the virtual system. Many system-level design considerations and best practices, including transcoding and various integrations with Cisco Unified Communications Manager, are explained in this section. In addition, this chapter provides details on third-party voicemail integration for supported industry-standard protocols.

This chapter presents a high-level design discussion and is focused on how the voice messaging products fit into a collaboration system with Unified CM. For detailed design guidelines for each product as well as interoperability information for third-party messaging and telephony systems, refer to the Cisco Unity Connection design guides, available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-implementation-design-guides-list.html

Voice Messaging Portfolio

The Cisco Unified Communications messaging portfolio consists of two main messaging products: Cisco Unity Connection and Cisco Unity Express. Each product fits different requirements yet each one contains overlapping features and scalability with regard to the others. They also have the ability to interwork with one another using Voice Mail Networking to achieve voicemail interoperability as well as higher scalability, as discussed later in this chapter.

When considering these products, it helps to think of the messaging types that the products apply to in order to understand the messaging options they include and to determine which options could fit your deployment requirements. The following definitions help describe these messaging types:

- *Voicemail-only* refers to a telephony voicemail integration where there is no access to the voicemail via any messaging client.
- *Integrated messaging* refers to voicemail with telephony access as well as voicemail-only access via a messaging client.
- *Unified messaging* refers to voicemail with telephony access as well as voicemail, email, and fax access via a messaging client.

Table 19-1 shows which Cisco products support these types of messaging.

Table 19-1 Supported Messaging Environments per Product

Messaging Type	Cisco Unity Connection	Cisco Unity Express
Voicemail-only	Yes	Yes
Integrated messaging	Yes	Yes
Unified messaging	Yes	No

Note For further details on Unified Messaging with Cisco Unity Connection, see Single Inbox with Cisco Unity Connection, page 19-44.

Based on the above messaging types and definitions, the two messaging product options are:

- Cisco Unity Connection

  This option combines unified and integrated messaging, voice recognition, and call transfer rules into an easy-to-manage system for medium-sized businesses with up to 20,000 users. In addition, multiple Cisco Unity Connection clusters can be joined using a digital or HTTPS network system. (Additionally, if required, two HTTPS or digital network systems can be joined using Voice Profile for Internet Mail (VPIM) networking to support more than 100,000 users.) Cisco Unity Connection can support up to 100,000 users in a digital network or HTTPS network. Cisco Unity Connection is also available with Cisco Business Edition. Cisco Business Edition 6000 supports up to a maximum

of 1,000 users. With Cisco Business Edition 7000, the normal Cisco Unity Connection capacity planning rules apply. For more information on Cisco Business Edition, see Design Considerations for Call Processing, page 9-26.

- Cisco Unity Express

  This option provides cost-effective voice and integrated messaging, automated attendant, and interactive voice response (IVR) capabilities in certain Cisco Integrated Services Routers (ISRs) for small and medium-sized businesses and enterprise branch offices with up to 500 users. When deployed as part of Cisco Business Edition 4000, Unity Express deployed as a container on Cisco ISR 4321 supports up to 200 users.

For a complete comparison of product feature, refer to the feature comparison documents available at

https://www.cisco.com/c/en/us/products/unified-communications/unity-connection/datasheet-listing.html

For more information on scalability of voice messaging products, refer to the section on Voice Messaging, page 25-42, in the chapter on Collaboration Solution Sizing Guidance, page 25-1.

This chapter focuses on the design aspects of integrating Cisco Unity Connection and Cisco Unity Express with Cisco Unified Communications Manager (Unified CM). Cisco Unified CM provides functionality for Session Initiation Protocol (SIP) trunks, which support integration directly to Cisco Unity Connection without the need for a SIP proxy server.

As mentioned, the design topics covered in this chapter apply to voicemail-only, unified messaging, and integrated messaging configurations. Additionally, this chapter discusses design aspects of deploying Cisco Unity Connection with Microsoft Exchange (2003, 2007, or 2010). Cisco Unity Connection and Unity Express have no dependencies on an external message store.

For additional design information about Cisco Unity Connection, including integrations with other non-Cisco messaging systems, refer to the latest version of the *Design Guide for Cisco Unity Connection*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-implementation-design-guides-list.html

For additional design information about Cisco Unity Express, including integrations with other non-Cisco messaging systems, refer to the applicable product documentation, available at

https://www.cisco.com/c/en/us/products/unified-communications/unity-express/index.html

Messaging Deployment Models

This section summarizes the various messaging deployment models for Cisco Unity Connection and Cisco Unity Express. For a complete discussion of the deployment models and design considerations specific to Cisco Unity Connection and the various messaging components, refer to the latest version of the *Design Guide for Cisco Unity Connection*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-implementation-design-guides-list.html

For Cisco Unity Express, refer to the applicable product documentation available at

https://www.cisco.com/c/en/us/products/unified-communications/unity-express/index.html

Cisco Unity Connection supports three primary messaging deployment models:

- Single-site messaging
- Multisite deployment with centralized messaging
- Multisite deployment with distributed messaging

Cisco Unity Express also supports three primary messaging deployment models:

- Single-site messaging
- Multisite deployment with distributed messaging
- Multisite deployment with distributed messaging with Cisco Unified CME

Note The Cisco Unity Express supports centralized voice messaging for up to 10 Unified CMEs. For more information, refer to the Cisco Unified Communications Manager Express documentation at https://www.cisco.com/c/en/us/products/unified-communications/unified-communications-manager-express/index.html.

Although the call processing deployment models for Cisco Unified CM and Unified CME are independent of the messaging deployment models for Cisco Unity Connection and Unity Express, each has implications toward the other that must be considered.

Cisco Unity Connection messaging redundancy is available in an active/active configuration. For more information, refer to the latest version of the *Design Guide for Cisco Unity Connection* available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-implementation-design-guides-list.html

All messaging deployment models support voicemail, integrated messaging, and unified messaging installations.

Single-Site Messaging

In this model, the messaging systems and messaging infrastructure components are all located at the same site, on the same highly available LAN. The site can be either a single site or a campus site interconnected via high-speed metropolitan area networks (MANs). All clients of the messaging system are also located at the single (or campus) site. The key distinguishing feature of this model is that there are no remote clients.

Centralized Messaging

In this model, similar to the single-site model, all the messaging system and messaging infrastructure components are located at the same site. The site can be one physical site or a campus site interconnected via high-speed MANs. However, unlike the single-site model, centralized messaging clients can be located both locally and remotely.

Distributed Messaging

A distributed messaging model consists of multiple single-site messaging systems distributed with a common messaging backbone. There can be multiple locations, each with its own messaging system and messaging infrastructure components. All client access is local to each messaging system, and the messaging systems share a messaging backbone that spans all locations. Message delivery from the distributed messaging systems occurs via the messaging backbone through a full-mesh or hub-and-spoke type of message routing infrastructure.

Distributed messaging is essentially multiple, single-site messaging models with a common messaging backbone. The exception to this rule is the PBX-IP Media Gateway (PIMG) and T1-IP Media Gateway (TIMG) integrations. PIMG and TIMG integrations are not discussed in this design document. For further information regarding PIMG or TIMG, refer to the latest Cisco Unity Connection integration guides available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-installation-and-configuration-guides-list.html

The distributed messaging model has the same design criteria as centralized messaging with regard to local and remote GUI clients, TRaP, and message downloads.

Messaging and Unified CM Deployment Model Combinations

This section discusses the design considerations for integrating the various messaging deployment models with the Unified CM call processing deployment models. Table 19-2 lists the various combinations of messaging and call processing deployment models supported by Cisco Unity Connection and Unity Express.

Table 19-2 Supported Combinations of Messaging and Unified CM Call Processing Deployment Models

Model Type	Cisco Unity Connection	Cisco Unity Express
Single-site messaging and single-site call processing	Yes	Yes
Centralized messaging and centralized call processing	Yes	No[1]
Distributed messaging and centralized call processing	Yes	Yes
Centralized messaging and distributed call processing	Yes	No[1]
Distributed messaging and distributed call processing	Yes	Yes
Centralized messaging with cluster over the WAN	Yes	No
Distributed messaging with cluster over the WAN	Yes	Yes

1. Support for centralized voicemail messaging with Unified CME is available with Cisco Unity Express; however, this is not applicable to Unified CM call processing deployment models.

This section covers the following topics:

- Cisco Unity Connection messaging and Unified CM deployment models
- Cisco Unity Express deployment models

Each topic defines a messaging and Unified CM deployment model combination and then highlights each Cisco voicemail messaging product applicable to that model as well as the design considerations for that model combination. Not all combinations are discussed for each product. Some examples are provided, with best practices and design considerations for each product. The intention is to provide an understanding of the base messaging deployment models and the interaction with Unified CM without detailing all possibilities.

For further details on site classification and a detailed analysis of supported combinations of messaging and call processing deployment models, refer to the latest version of the *Design Guide for Cisco Unity Connection*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-implementation-design-guides-list.html

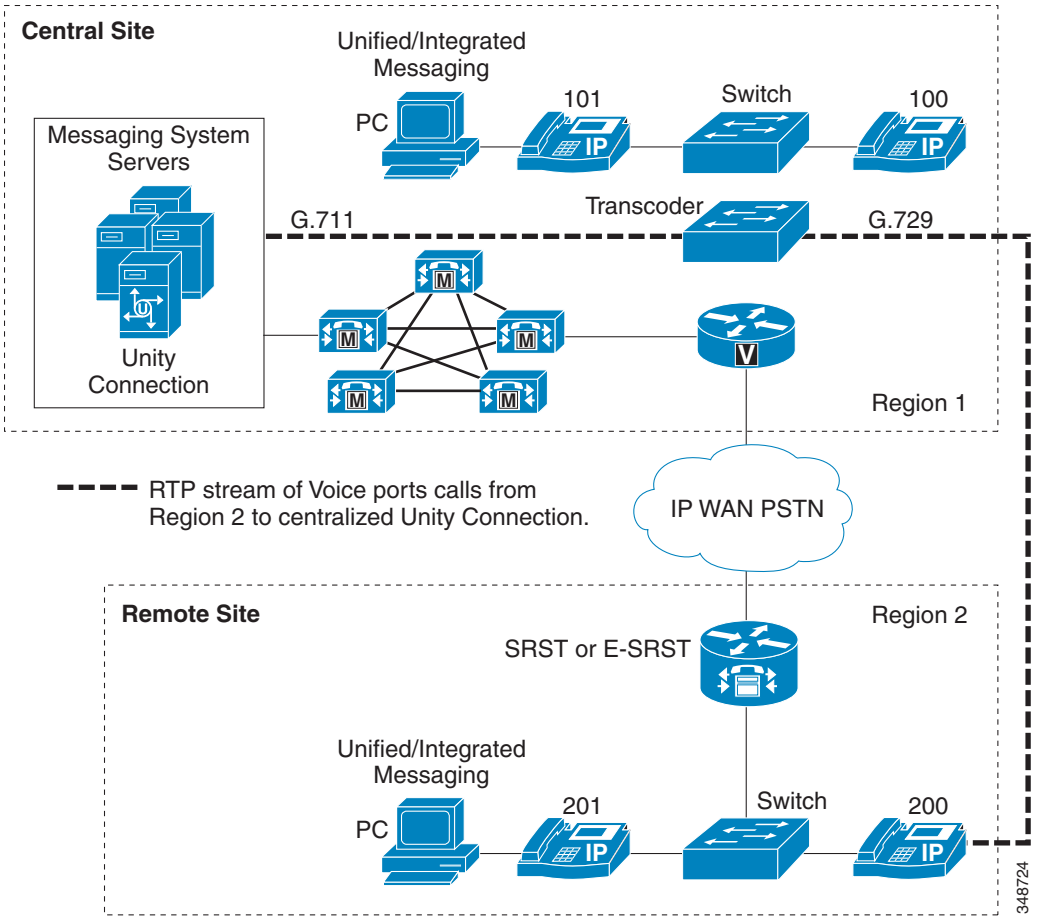
Cisco Unity Connection Messaging and Unified CM Deployment Models

This section discusses some of the various combinations of messaging and call processing deployment models for Cisco Unity Connection.

Centralized Messaging and Centralized Call Processing

In centralized messaging, the voice messaging server is located in the same site as the Unified CM cluster. With centralized call processing, subscribers may be located either remotely and/or locally to the cluster and messaging server(s). (See Figure 19-1.) When remote users access resources at the central site (such as voice ports, IP phones, or PSTN gateways, as in Tail-End Hop-Off (TEHO)), these calls are transparent to gatekeeper call admission control. Therefore, regions and locations must be configured in Unified CM for call admission control. (See Managing Bandwidth, page 19-32.) When making inter-region calls to IP phones or MGCP gateways, IP phones automatically select the inter-region codec that has been configured.

Figure 19-1 Centralized Messaging with Centralized Call Processing



In Figure 19-1, regions 1 and 2 are configured to use G.711 for intra-region calls and G.729 for inter-region calls.

As Figure 19-1 shows, when a call is made from extension 200 to the voicemail ports in Region 1, the inter-region G.729 codec is used at the endpoint but the RTP stream is transcoded to use G.711 on the voice ports. Unified CM transcoding resources must be located at the same site as the voicemail system.

Impact of Non-Delivery of RDNIS on Voicemail Calls Routed by AAR

In centralized messaging environments, automated alternate routing (AAR), a Unified CM feature, can route calls over the PSTN to the messaging store at the central site when the WAN is oversubscribed. However, when calls are rerouted over the PSTN, Redirected Dialed Number Information Service (RDNIS) can be affected. Incorrect RDNIS information can impact voicemail calls that are rerouted over the PSTN by AAR when Cisco Unity Connection is remote from its messaging clients. If the RDNIS information is not correct, the call will not reach the voicemail box of the dialed user but will instead receive the auto-attendant prompt, and the caller might be asked to re-enter the extension number of the party they wish to reach. This behavior is primarily an issue when the telephone carrier is unable to ensure RDNIS across the network. There are numerous reasons why the carrier might not be able to ensure that RDNIS is properly sent. Check with your carrier to determine if they provide guaranteed RDNIS deliver end-to-end for your circuits. The alternative to using AAR for oversubscribed WANs is simply to let callers hear reorder tone in an oversubscribed condition.

Cisco Unity Connection Survivable Remote Site Voicemail

Cisco Unity Connection Survivable Remote Site Voicemail (SRSV) is used in the centralized Cisco Unified Communications Manager and Cisco Unity Connection deployment model that provides survivability of voicemail service for branch site users in the event of WAN outage. Cisco Unity Connection now provides the SRSV functionality natively without using a Cisco Unified Messaging Gateway at the central site. Cisco Unity Connection SRSV is a replacement option for Cisco Unity Express SRSV. During normal operation Cisco Unity Connection updates the information about phones and user mailboxes to the branch site SRSV server. In the event of a WAN outage the Unity Connection SRSV branch server acts as a backup auto-attendant and voicemail storage. All incoming unanswered and busy calls are forwarded to the Unity Connection SRSV branch server, where external or internal callers may leave voice messages.

Upon WAN restoration, all the voicemail is deleted from the branch site SRSV and uploaded to the central Cisco Unity Connection server. Once the upload is complete, the branch site SRSV moves to an idle state. All the incoming unanswered and busy calls are again forwarded to the central Cisco Unity Connection server.

SRSV Deployment Models

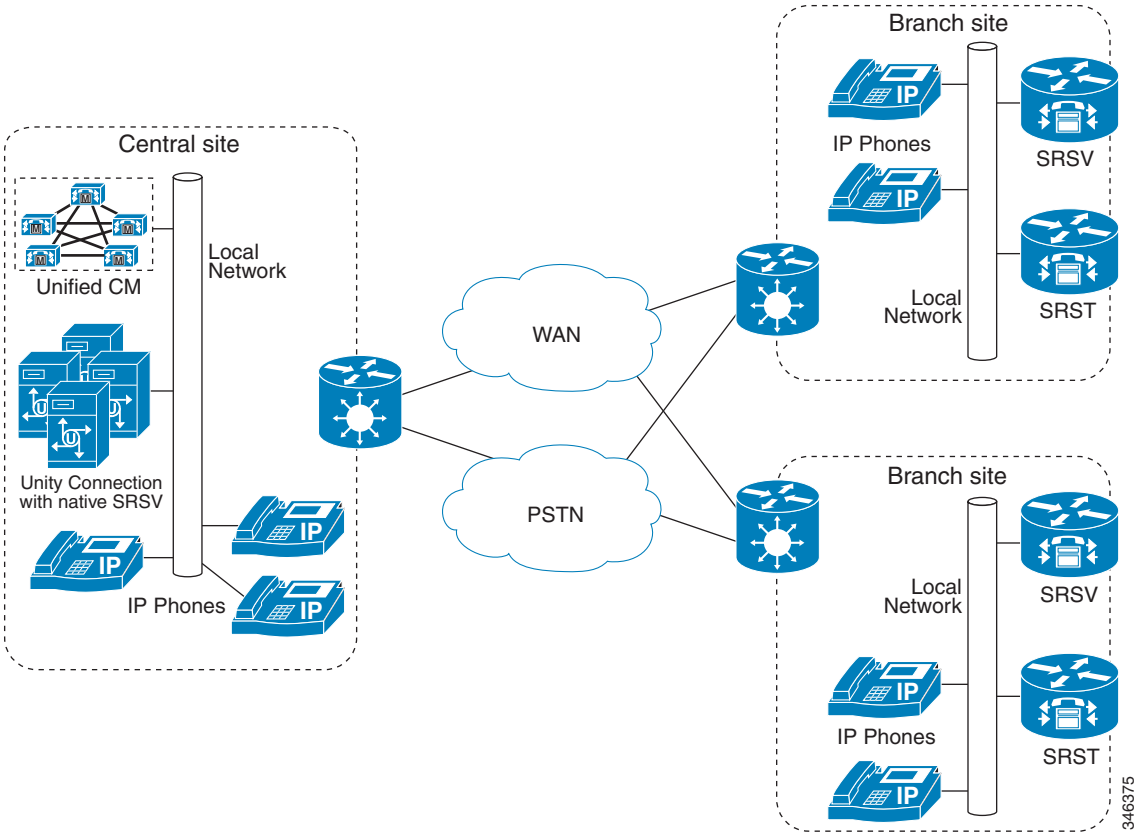
The following deployment models support Survivable Remote Site Voicemail (SRSV):

- SRST or E-SRST at the Branch Site with Centralized Unified CM and Unity Connection, page 19-8
- Multiple E-SRST or SRST Servers at the Branch Site with Centralized Unified CM and Unity Connection, page 19-9

SRST or E-SRST at the Branch Site with Centralized Unified CM and Unity Connection

As shown in Figure 19-2, the central site contains Cisco Unified CM and Unity Connection to provide primary call processing and voice messaging services under normal conditions. At the branch site, Cisco Unified Enhanced Survivable Remote Site Telephony (E-SRST) and a Cisco Unity Connection SRSV branch server are installed as a backup call agent and voice messaging server in the event of WAN outage. Cisco Unity Connection installed at the central site uploads all phone and voice mailbox information to the branch site SRSV server. SRST or E-SRST remains in the idle state until connectivity to central site is lost. Once the branch site becomes isolated from central site and the keep-alive timer between phones and Unified CM expires, branch phones are re-homed to the E-SRST or SRST router which is preconfigured to send unanswered and busy calls to the Unity Connection SRSV branch server. Subscribers can listen to voice messages left during a WAN outage by accessing voicemail. Upon WAN restoration, all the voice messages are uploaded to a subscriber mailbox on the central Cisco Unity Connection.

Figure 19-2 SRST or E-SRST at Branch Site with Centralized Unified CM and Unity Connection

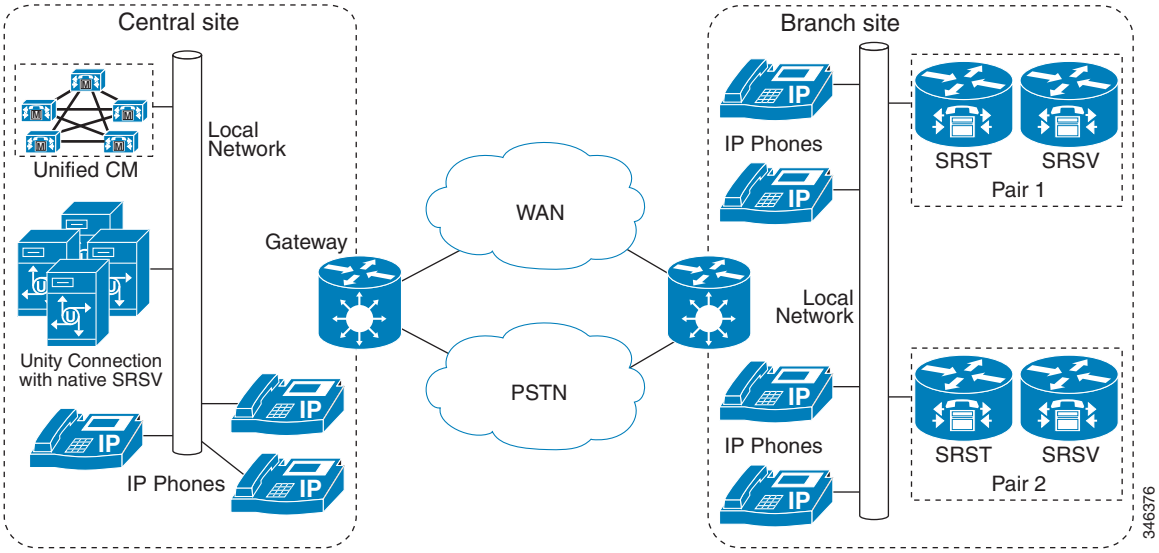


Multiple E-SRST or SRST Servers at the Branch Site with Centralized Unified CM and Unity Connection

This deployment model is similar to first scenario, but multiple E-SRST and Cisco Unity Connection SRSV branch servers are paired at the branch site for load balancing (see Figure 19-3). The administrator must manually divide branch site users across two E-SRST servers using two different SRST references in Unified CM to achieve load balancing. Cisco Unity Connection pushes the mailbox information to the appropriate paired Cisco Unity Connection SRSV branch server. With this configuration, each Cisco Unity Connection SRSV branch server contains the mailboxes for users on a single branch E-SRST.

Each Cisco Unity Connection SRSV branch server handles calls forwarded from its paired E-SRST router in the event of WAN outage. Similar to the first scenario, the Cisco Unity Connection SRSV branch server uploads all voicemail to a subscriber mailbox in the central Cisco Unity Connection upon WAN restoration.

Figure 19-3 Multiple E-SRST or SRST Servers at Branch Site with Centralized Unified CM and Unity Connection



Note Pairing a single Cisco Unity Connection SRSV branch server with multiple E-SRST servers at a branch site is not supported.

Deployment Guidelines for Survivable Remote Site Voicemail

- The maximum number of supported remote sites is 35 per central Cisco Unity Connection. For more information on supported remote sites with each virtual platform overlay, refer the latest version of the *Cisco Unity Connection Supported Platforms List*, available at

  https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-installation-guides-list.html

- This solution supports both fallback methods, SRST and Enhanced SRST (E-SRST). The Cisco Unity Connection SRSV branch server runs on the Cisco Services-Ready Engine (SRE) 900 and 910 blade servers and any supported Cisco Unity Connection platform such as Cisco Unified Computing System (UCS) or UCS E-Series. Both the Cisco Unity Connection SRSV branch server and the SRST or E-SRST router appear as a single logical unit, where the SRST router handles all control signaling in the event the of a WAN outage.
- The Cisco Unity Connection SRSV branch server becomes active if the WAN link goes down and SRST is in active state. Otherwise it remains in the idle state.
- Use HTTP over Secure Socket Layer (SSL) protocol to secure the connection between Cisco Unity Connection and Cisco Unity Connection SRSV.

SRSV uses bandwidth from the WAN link during the following activities:

- Configuration uploads from Cisco Unity Connection to Cisco Unity Connection SRSV
- Uploading of voice messages from the branch Cisco Unity Connection SRSV server to the central Cisco Unity Connection when the WAN link is restored

Distributed Messaging with Centralized Call Processing

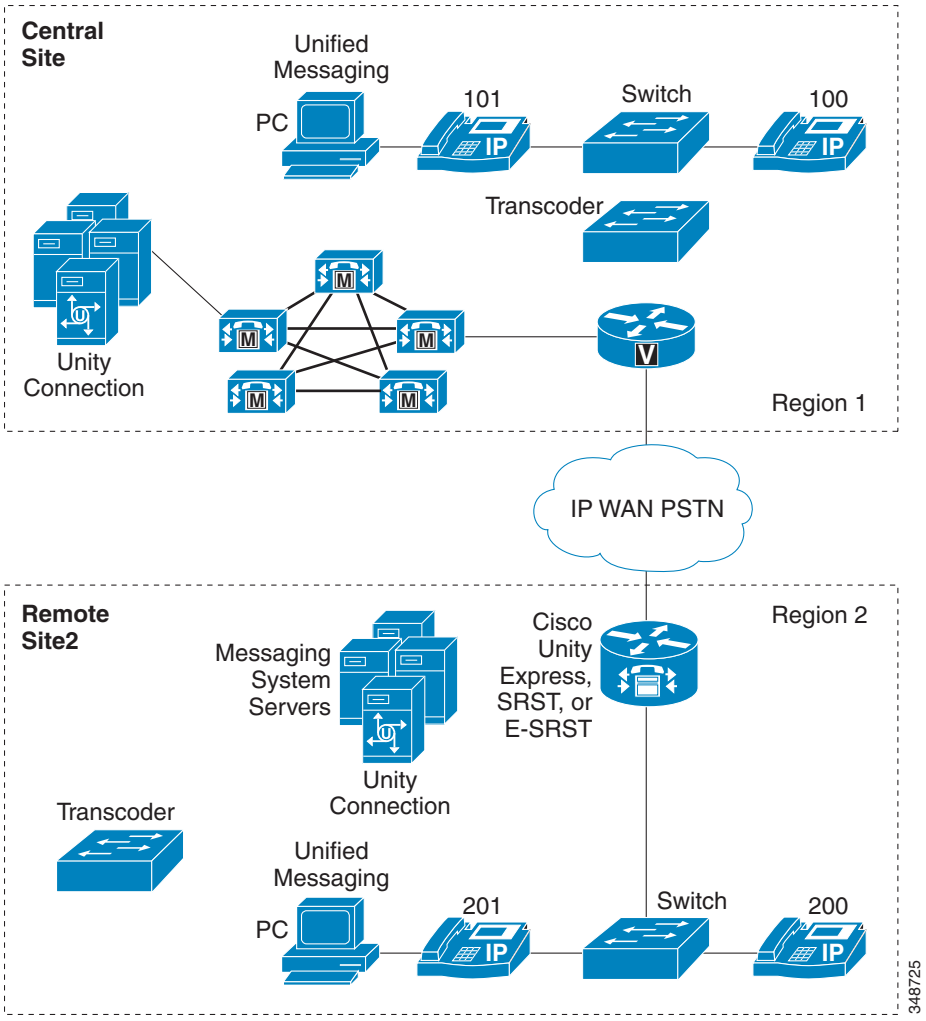
Distributed messaging means that there are multiple messaging systems distributed within the telephony environment, and each messaging system services only local messaging clients. This model differs from centralized messaging, where clients are both local and remote from the messaging system.

Figure 19-4 illustrates the distributed messaging model with centralized call processing. As with other multisite call processing models, the use of regions and locations is required to manage WAN bandwidth.

Note that Cisco Unified Communications Manager Express in E-SRST mode is used for call processing backup of both IP phones and Cisco Unity Connection voicemail ports. Deployed at the remote site (for example, Region 2 in Figure 19-4), this fallback support provides backup call processing in the event that the phones lose connectivity with Unified CM, such as during a WAN failure, while simultaneously providing users at the remote site with access to the local Cisco Unity Connection server as well as MWI support during WAN failure. For further details on E-SRST mode, refer to the documentation at

https://www.cisco.com/c/en/us/products/unified-communications/unified-survivable-remote-site-telephony/index.html

Figure 19-4 Distributed Messaging with Centralized Call Processing



For the configuration in Figure 19-4, transcoder resources must be local to each Cisco Unity Connection message system site. Regions 1 and 2 are configured to use G.711 for intra-region calls and G.729 for inter-region calls.

Voice messaging ports for both Cisco Unity Connection servers must be assigned the appropriate region and location by means of calling search spaces and device pools configured on the Unified CM server. In addition, to associate telephony users with a specific group of voicemail ports, you must configure Unified CM voicemail profiles. For details on configuring calling search spaces, device pools, and voicemail profiles, refer to the applicable version of the *Administration Guide for Cisco Unified Communications Manager and IM and Presence Service*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

Cisco Unity Connection supports digital and HTTPS networking, which enables multiple Unity Connection clusters deployed over a WAN to communicate with each other. Using digital or HTTPS networking, multiple Unity Connection clusters can share common directory information. This allows users on multiple clusters to leave voicemail to each other. The Cisco Unity Connection cluster can integrate with a corporate directory such as Microsoft Active Directory to synchronize user information and can use digital or HTTPS networking to share the directory information at the same time.

Cisco Unity Connection with E-SRST

E-SRST offers the possibility for Cisco Unity Connection servers located in remote sites and registered with a Unified CM at the central site to fall-back to E-SRST in the remote location. When the WAN link is down and the phones fail-over to the E-SRST router, Cisco Unity Connection voicemail ports can also fail-over to E-SRST mode to provide the remote site users with access to their voicemail with MWI during the WAN outage.

Note MWI has to be resynchronized from the Cisco Unity Connection server whenever a failover happens from Unified CM to E-SRST mode, or vice versa.

Combined Messaging Deployment Models

It is possible to combine messaging models in the same deployment, provided that the deployment adheres to all the guidelines listed in the preceding sections. Figure 19-5 shows a user environment in which both centralized and distributed messaging are employed simultaneously.

Figure 19-5 Combined Deployment Models

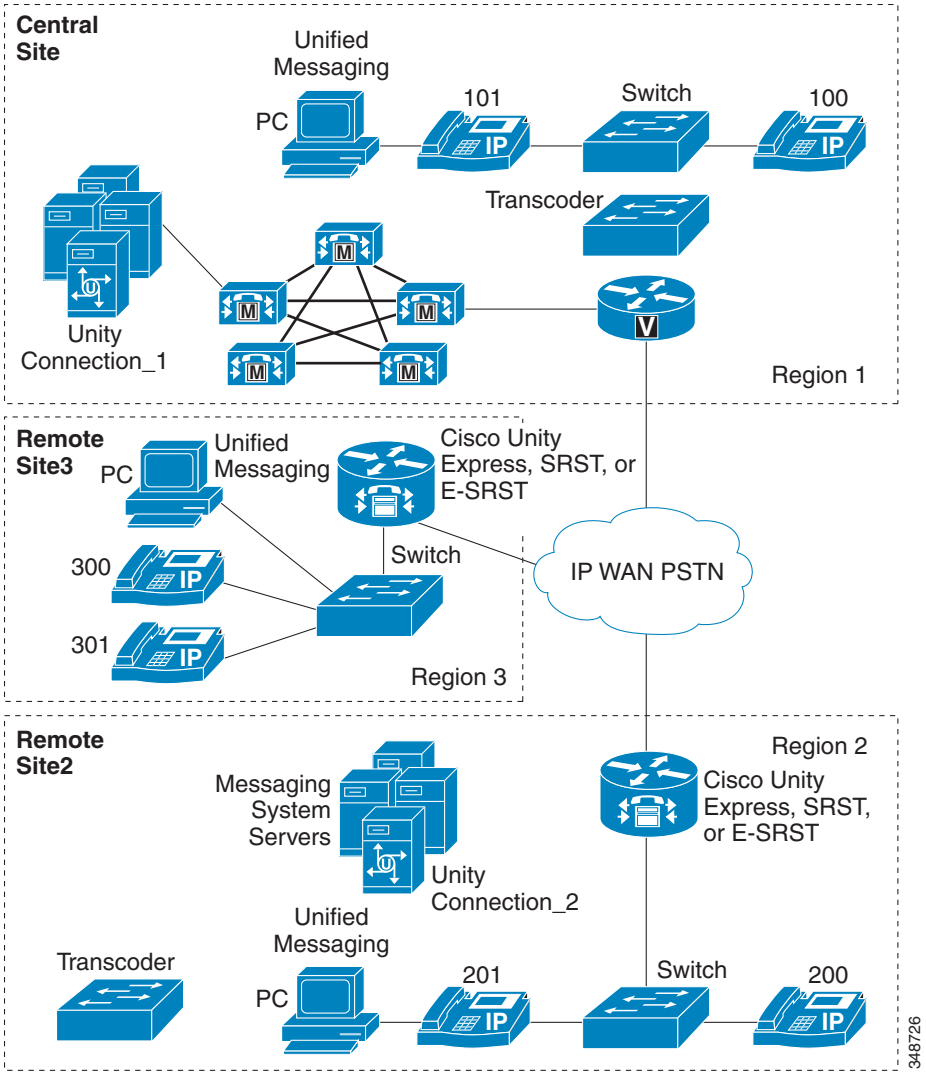


Figure 19-5 shows the combination of two messaging models. Regions 1 and 3 use centralized messaging with centralized call processing, while Region 2 uses distributed messaging with centralized call processing. All regions are configured to use G.711 for intra-region calls and G.729 for inter-region calls.

In Figure 19-5, centralized messaging and centralized call signaling are used between the Central Site and Site3. The messaging system at the Central Site provides messaging services for clients at both the Central Site and Site3. Site2 uses the distributed messaging model with centralized call processing. The messaging system (Unity Connection 2) located at Site2 provides messaging services for only those

users located within Site2. In this deployment, both models adhere to their respective design guidelines as presented in this chapter. Transcoding resources are located locally to each messaging system site, and they support clients who access messaging services from a remote site (relative to the messaging system), as in the case of a Site2 user leaving a message for a Central Site user.

In addition, E-SRST mode is used for call processing backup of both IP phones and Cisco Unity Connection voicemail ports. Deployed at the remote site (for example, Region 2 in Figure 19-5), this fallback support provides backup call processing in the event that the phones lose connectivity with Unified CM, such as during a WAN failure, while simultaneously providing users at the remote site with access to the local Cisco Unity Connection server as well as MWI support during WAN failure. For further details on E-SRST, refer to the product documentation available at

https://www.cisco.com/c/en/us/products/unified-communications/unified-survivable-remote-site-telephony/index.html

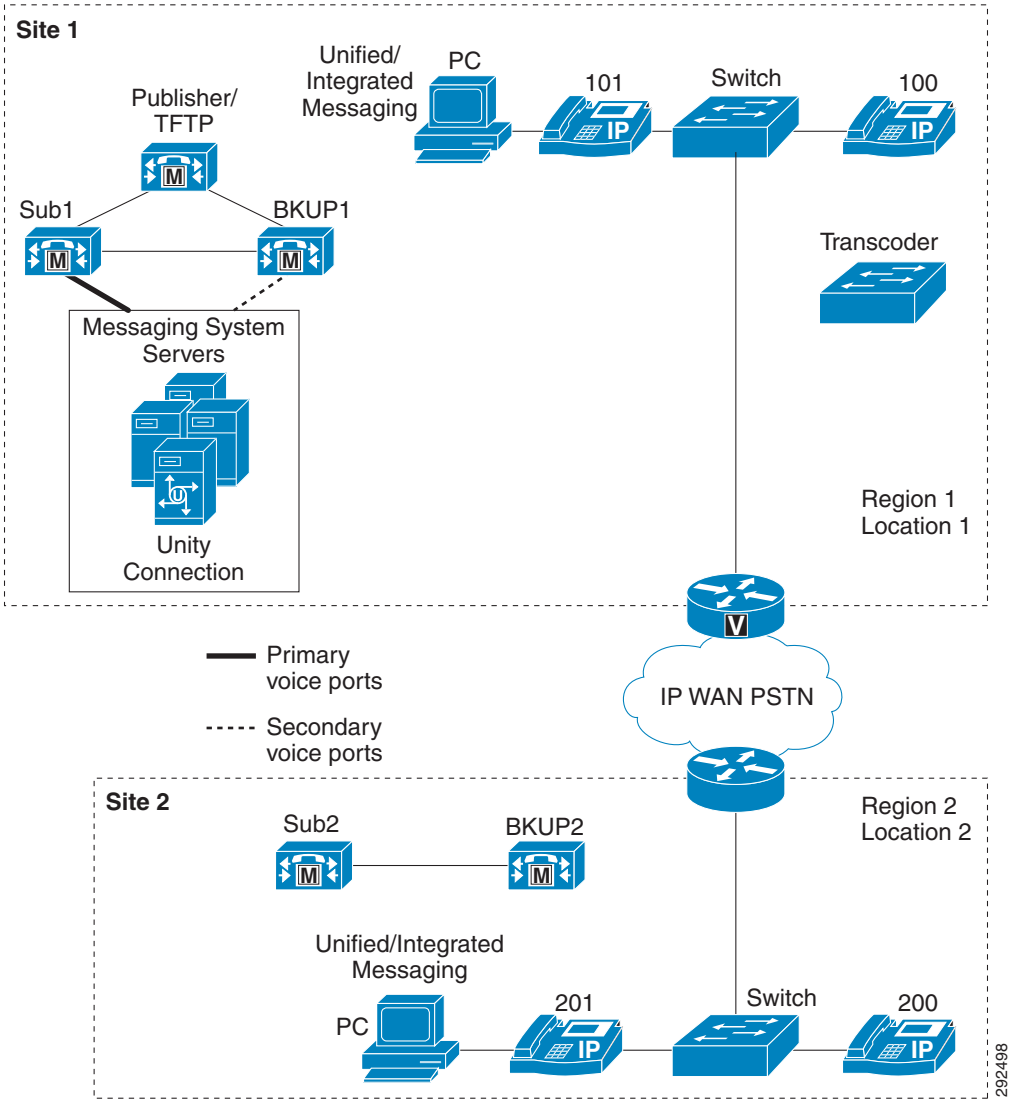
Centralized Messaging with Clustering Over the WAN

This section addresses Cisco Unity Connection design issues for deploying centralized messaging with Unified CM clustering over the WAN with local failover. In the case of a WAN failure with this model, all remote messaging sites will lose voicemail capability until the WAN is restored. (See Figure 19-6.)

Clustering over the WAN supports local failover. With local failover, each site has a backup subscriber server physically located at the site. This section focuses on deploying Cisco Unity Connection centralized messaging with local failover for clustering over the WAN.

For additional information, refer to the section on Clustering Over the IP WAN, page 10-43.

Figure 19-6 Cisco Unity Connection Centralized Messaging and Clustering Over the WAN with Local Failover



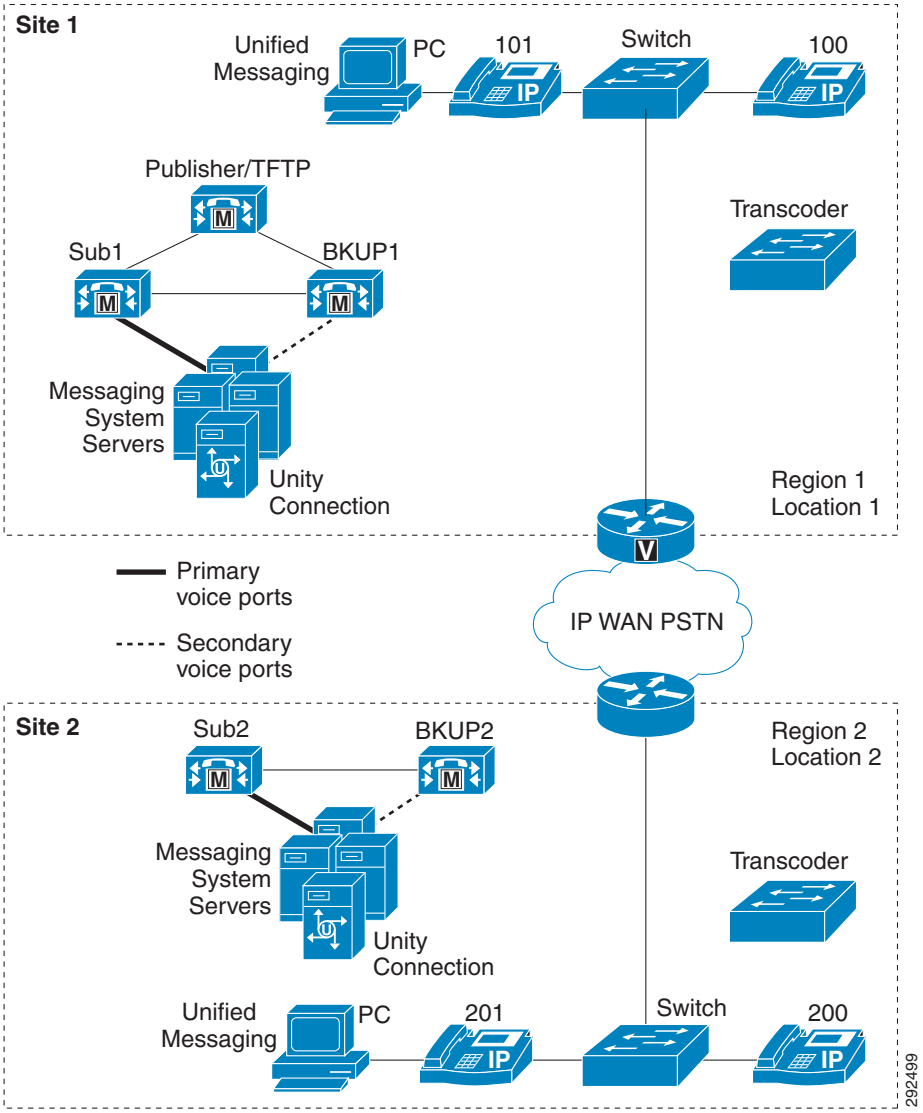
For minimum bandwidth requirements between clustered servers see the section on Local Failover Deployment Model, page 10-47.

Clustering over the WAN with Unified CM supports up to eight sites, as does Cisco Unity Connection. The voicemail ports are configured only at the site where the Cisco Unity Connection messaging system is located (see Figure 19-6). Voicemail ports do not register over the WAN to the remote site(s). Messaging clients at the other site(s) access all voicemail resources from the primary site. There is no benefit to configuring voice ports over the WAN to any of the remote sites because, in the event of a WAN failure, remote sites would lose access to the centralized messaging system. Because of bandwidth consideration, the voicemail ports should have TRaP disabled and all messaging clients should download voicemail messages to their local PCs (unified messaging only).

Distributed Messaging with Clustering Over the WAN

Local failover sites that also have Cisco Unity Connection messaging server(s) deployed would have voice ports registered to the local Unified CM subscriber server(s), similar to the centralized messaging model. For information about configuring the voice ports, see Voice Port Integration with a Unified CM Cluster, page 19-40, and Voice Port Integration with Dedicated Unified CM Backup Servers, page 19-42.

Figure 19-7 Cisco Unity Connection Distributed Messaging and Clustering over the WAN



In a purely distributed messaging implementation with clustering over the WAN, each site in the cluster would have its own Cisco Unity Connection messaging server with messaging infrastructure components. If not all of the sites have local Cisco Unity Connection messaging systems but some sites have local messaging clients using a remote messaging server(s), this deployment would be a combination model with both distributed messaging and centralized messaging. (See Combined Messaging Deployment Models, page 19-13.) In the event of a WAN failure in this model, all remote sites that use centralized messaging will lose voicemail capability until the WAN is restored.

Each site that does not have a local messaging server must use a single messaging server for all of its messaging clients, but all such sites do not have to use the same messaging server. For example, suppose Site1 and Site2 each have a local messaging server. Site3 can then have all of its clients use (register with) the messaging server at Site2, while Site4 can have all of its clients use the messaging server at Site1. Transcoder resources are required at sites that have local Cisco Unity Connection messaging server(s).

As with other distributed call processing deployments, calls going between these sites are transparent to gatekeeper call admission control, therefore you must configure regions and locations in Unified CM to provide call admission control. (See Managing Bandwidth, page 19-32.)

The distributed Cisco Unity Connection servers may also be networked using digital or HTTPS networking.

Messaging Redundancy

Messaging redundancy is discussed in this section as it refers to Cisco Unity Connection. Cisco Unity Express does not support messaging redundancy.

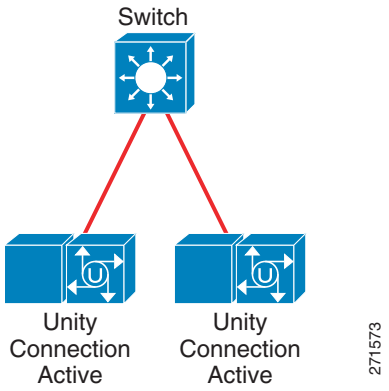
Cisco Unity Connection

Cisco Unity Connection supports messaging redundancy and load balancing in an active-active redundancy model consisting of two servers, a primary and a secondary, configured as an active/active redundant pair of servers, where both the primary and secondary servers actively accept calls as well as HTTP and IMAP requests. For more information, refer to the latest version of the *Design Guide for Cisco Unity Connection*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-implementation-design-guides-list.html

Figure 19-8 illustrates Cisco Unity Connection active/active messaging redundancy.

Figure 19-8 Redundancy of Cisco Unity Connection Messaging



Cisco Unity Connection SIP trunk implementation requires call forking for messaging redundancy functionality. Cisco Unified Communications Manager supports the multi-destination SIP trunk feature. With this multi-destination SIP trunk feature, administrators can define full-mesh trunking between Cisco Unified CM and Cisco Unity Connection to achieve redundancy. Also, two separate SIP trunks

can be configured, one for each server in a pair, and they can be added to the same route group associated to the same route list.The route group should be configured in top-down order so that calls are sent to the primary Unity Connection and overflow calls are sent to secondary Unity Connection server.

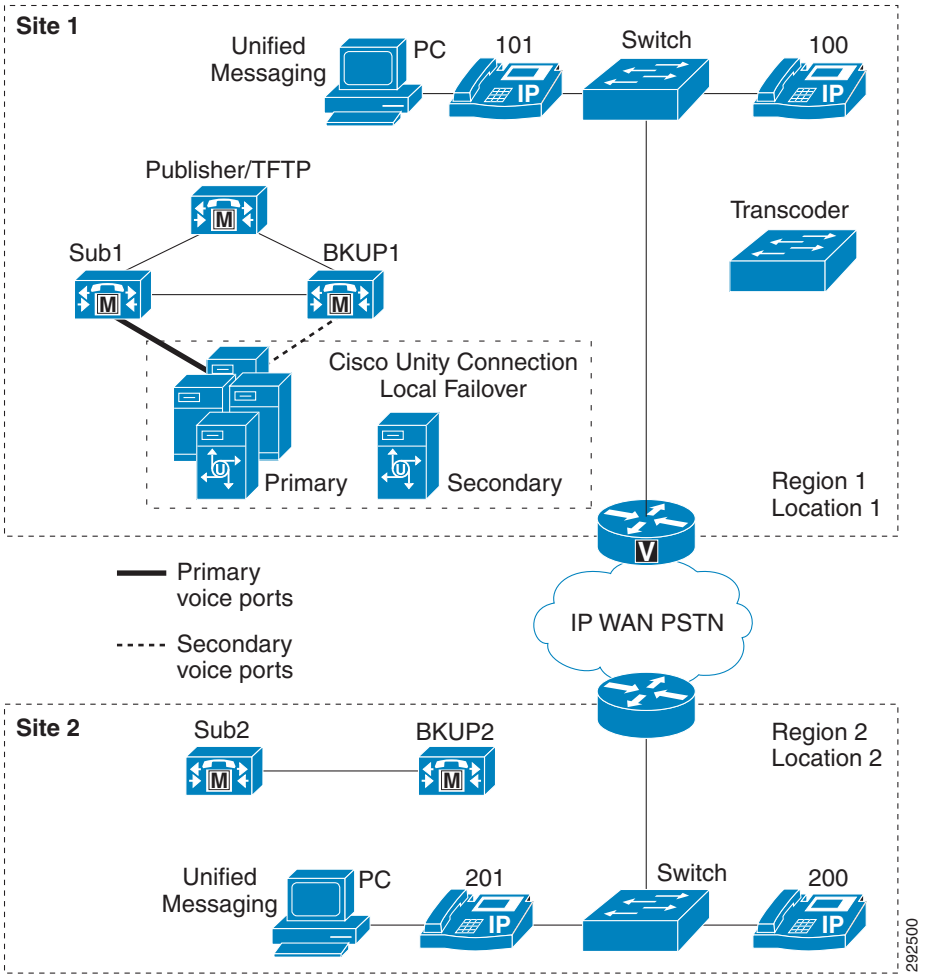
Note SIP OPTIONS Ping should be enabled on the Cisco Unified CM SIP trunk for Cisco Unity Connection failover to work properly.

Cisco Unity Connection Failover and Clustering Over the WAN

When deploying Cisco Unity Connection local failover with clustering over the WAN, apply the same design practices described in Centralized Messaging with Clustering Over the WAN, page 19-14, and Distributed Messaging with Clustering Over the WAN, page 19-16. The voice ports from the primary Cisco Unity Connection server should not cross the WAN during normal operation.

Figure 19-9 depicts Cisco Unity Connection local failover. Note that the primary and secondary Cisco Unity Connection servers are both physically located at the same site. Cisco Unity Connection failover supports up to the maximum number of remote sites available with clustering over the WAN for Unified CM.

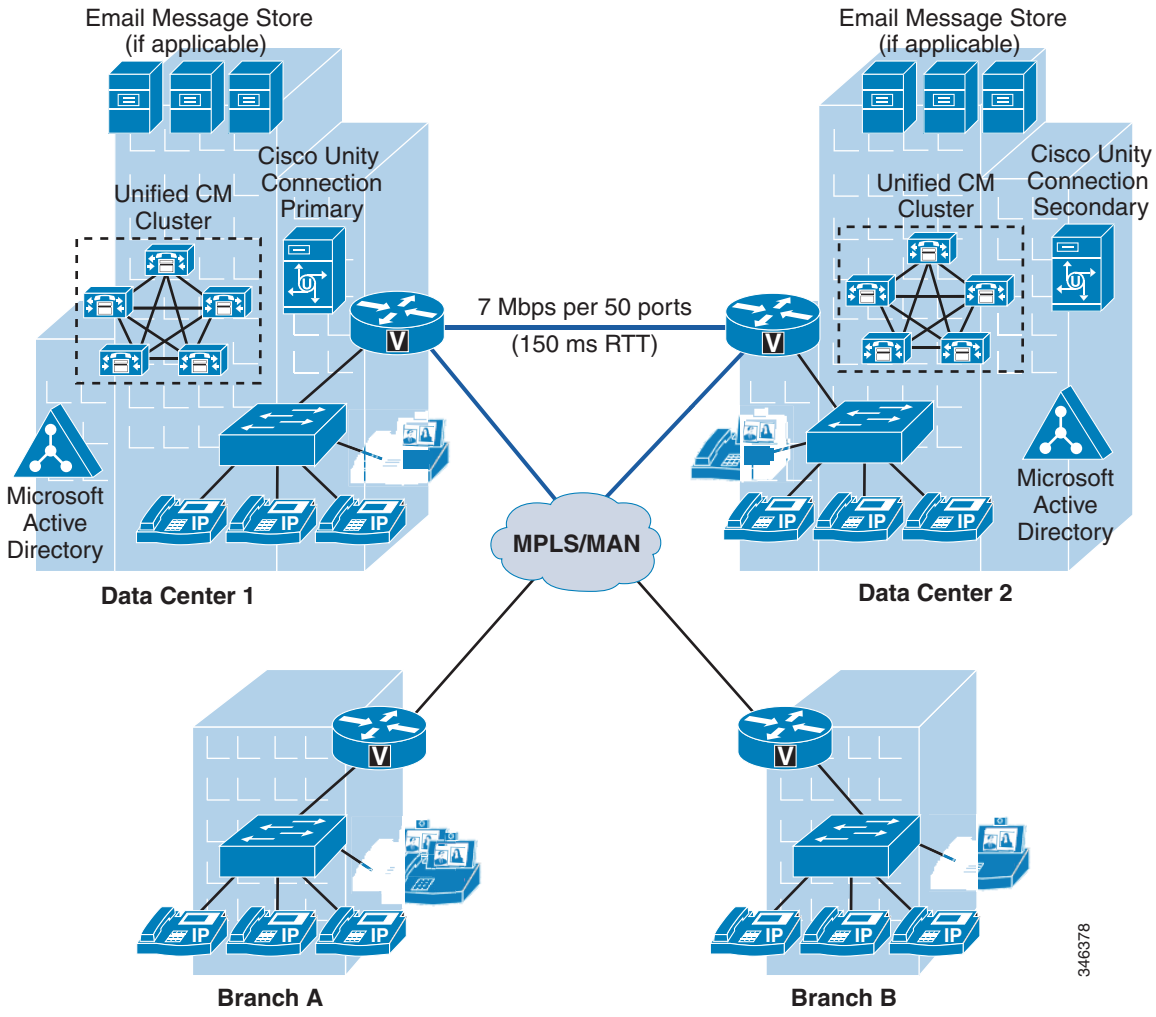
Figure 19-9 Cisco Unity Connection Local Failover and Clustering Over the WAN



Cisco Unity Connection Redundancy and Clustering Over the WAN

Cisco Unity Connection supports both active/active and active/standby clustering for redundancy and can be deployed over the WAN. The active/active or "high availability" configuration provides both high availability and redundancy. Both servers in the active/active pair run the Cisco Unity Connection application to accept calls as well as HTTP and IMAP requests from clients. The Cisco Unity Connection primary server handles all the incoming calls and administrative changes in an active/standby deployment. The only time the secondary server would handle calls in this scenario is when the primary server is in a failed state or unavailable. Each of the servers from the cluster can be deployed over the WAN at different sites, following the required design consideration. Figure 19-10 depicts a Cisco Unity Connection deployment with clustering over WAN for geographically separated data centers.

Figure 19-10 Cisco Unity Connection with High Availability Between Two Sites



Consider the following delay and bandwidth requirements when deploying Cisco Unity Connection servers over different sites:

- Maximum of 100 ms RTT between an active/active pair at different sites.
- Maximum of 150 ms RTT between an active/standby pair at different sites.
- Minimum of 7 Mbps bandwidth is required for every 50 ports. (For example, 250 ports require 35 Mbps.)

Note Bandwidth and latency requirements may differ for different versions of Cisco Unity Connection.

For a complete set of requirements, refer to the latest version of the *System Requirements for Cisco Unity Connection*, available at

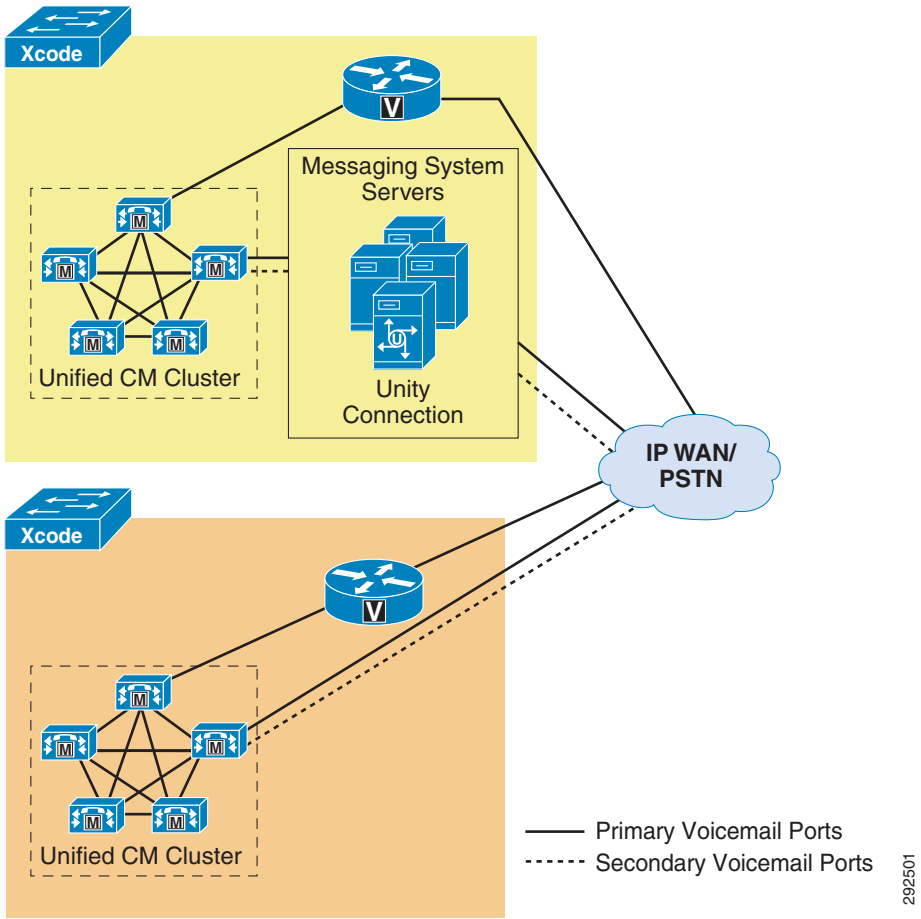
https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-installation-guides-list.html

Note The Cisco Unity Connection cluster feature is also supported with Cisco Business Edition 6000.

Centralized Messaging with Distributed Unified CM Clusters

Cisco Unity Connection can also be deployed in a centralized messaging configuration with multiple Unified CM clusters (see Figure 19-11). See the section on Integration with Cisco Unified CM, page 19-35, for details on multiple integrations and MWI considerations with multiple Unified CM clusters.

Figure 19-11 Integrating Cisco Unity Connection with Multiple Unified CM Clusters



For the configuration in Figure 19-11, messaging clients at both Cluster 1 and Cluster 2 sites use the Cisco Unity Connection messaging infrastructure physically located at Cluster 1.

Cisco Unity Express Deployment Models

This section begins with a quick overview of Cisco Unity Express, covering product related information. Next the deployment models section presents three supported deployment models with Cisco Unity Express, focusing on distributed voice messaging with both centralized and distributed call processing followed by some deployment characteristics and design guidelines. Lastly, this section discusses the signaling call flows and the various protocols used between Cisco Unity Express and Unified CM as well as between Cisco Unity Express and Unified SRST or E-SRST mode.

Overview of Cisco Unity Express

Cisco Unity Express is Linux-based software running on a Cisco Network Module in Cisco Integrated Services Routers (ISRs). It is an entry-level auto-attendant (AA) and voicemail solution that can be deployed with Cisco Unified Communications Manager (Unified CM), Cisco Unified SRST, or Cisco Unified Communications Manager Express (Unified CME). In prior releases, Cisco Unity Express was limited to a co-resident deployment with Unified CME or a Survivable Remote Site Telephony (SRST) router. However, with the H.323-to-SIP call routing capability introduced in Cisco IOS Release 12.3(11)T, Cisco Unity Express and SRST or Unified CME can reside on two separate routers when deployed with Unified CM or Unified CME, respectively. Cisco Unity Express uses SIP to communicate with Cisco Unified Communications Manager Express (Unified CME) while Cisco Unity Express uses JTAPI to connect to Cisco Unified Communications Manager (Unified CM).

For more information on supported hardware platforms and capacity with Cisco Unity Express, refer to the product release note available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-express/products-release-notes-list.html

For details on interoperability of Unified CM and Unified CME, see Integration of Multiple Call Processing Agents, page 9-36.

For additional information on supported deployment models with Unified CME, refer to the appropriate Cisco Unified Communications Manager Express design documentation available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-express/products-implementation-design-guides-list.html

Deployment Models

Cisco Unity Express can be deployed as a single site or distributed voicemail and automated attendant (AA) solution for Cisco Unified Communications Manager (Unified CM) or Unified Communications Manager Express (Unified CME). However, Cisco Unity Express is supported with all of the Cisco Unified CM deployment models, including:

- Single-site deployments
- Multisite deployments with centralized call processing
- Multisite deployments with distributed call processing

Figure 19-12 shows a centralized call processing deployment incorporating Cisco Unity Express, and Figure 19-13 shows a distributed call processing deployment.

Cisco Unity Express sites controlled by Unified CME, as well as other sites controlled by Unified CM, can be interconnected with each other using SIP trunking protocol. Although Cisco Unity Express can integrate with either Unified CM or Unified CME, it cannot integrate with both simultaneously.

Note Cisco Unity Express supports a centralized deployment model with up to 10 Unified CMEs.

Figure 19-12 Cisco Unity Express in a Centralized Call Processing Deployment

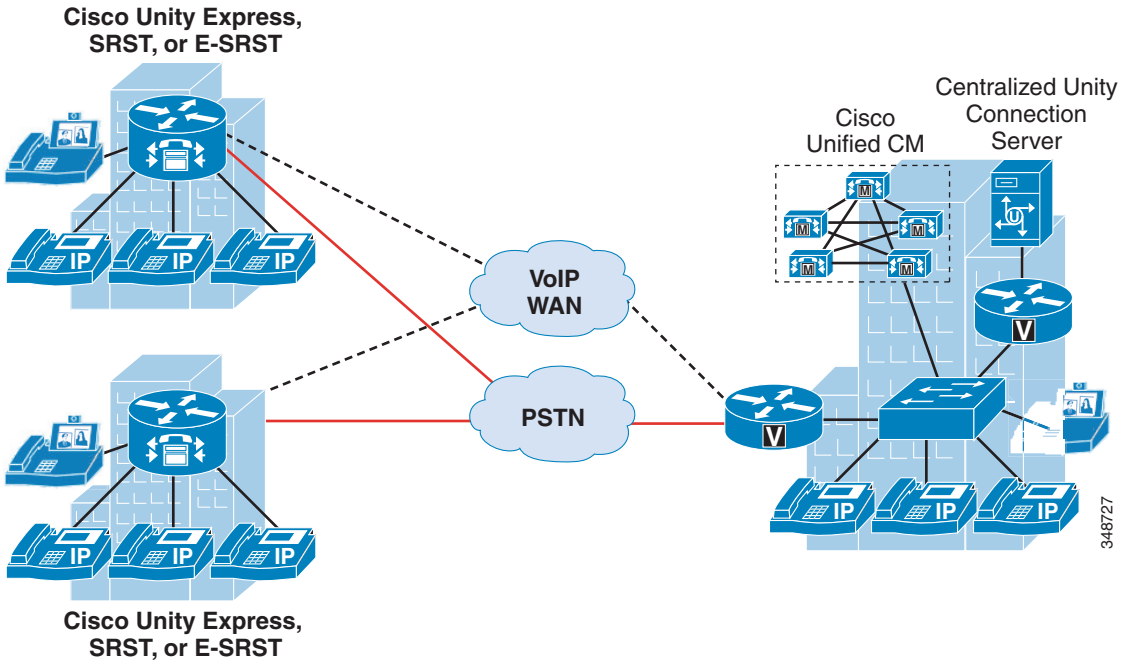
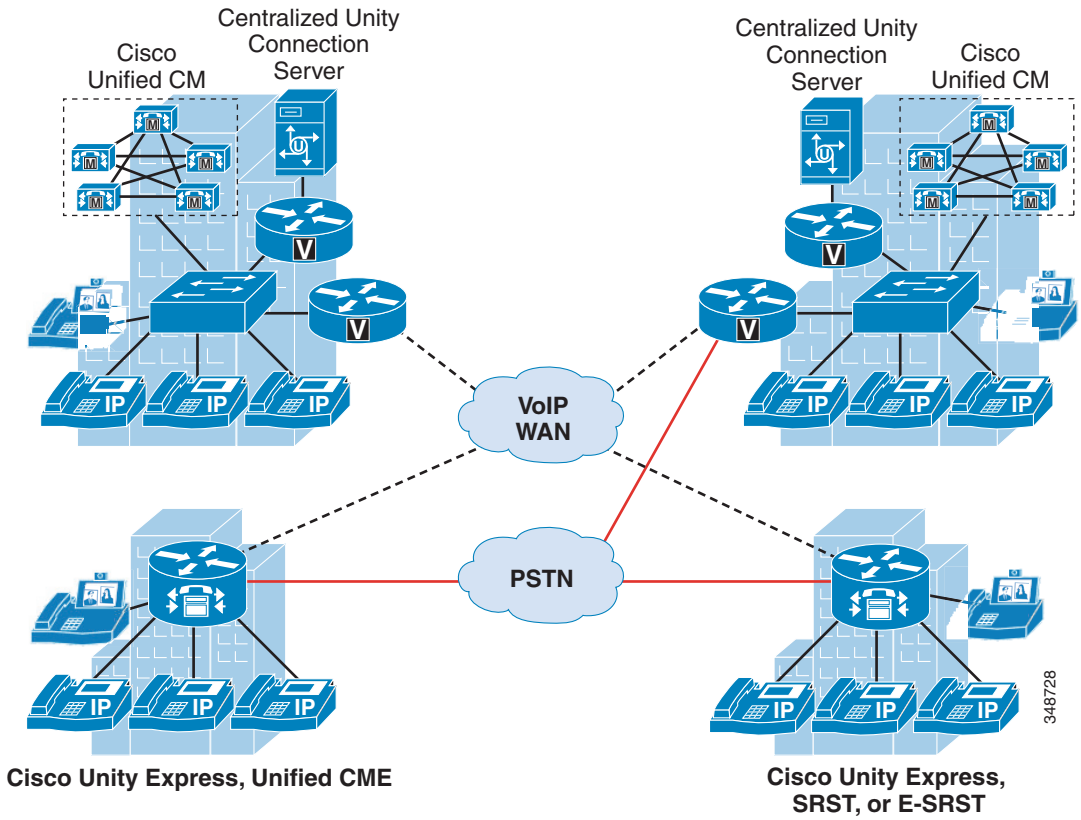


Figure 19-13 Cisco Unity Express in a Distributed Call Processing Deployment



The most likely deployment model to use Cisco Unity Express is the multisite WAN model with centralized call processing, where Cisco Unity Express provides distributed voicemail at the smaller remote offices and a central Cisco Unity Connection system provides voicemail to the main campus and larger remote sites.

Use Cisco Unity Express as a distributed voicemail solution if any of the following conditions apply to your Unified CM network deployment:

- Survivability of voicemail and AA access must be ensured regardless of WAN availability.
- Available WAN bandwidth is insufficient to support voicemail calls traversing the WAN to a central voicemail server.
- There is limited geographic coverage of the AA or branch site PSTN phone numbers published to the local community, and these numbers cannot be dialed to reach a central AA server without incurring toll charges.
- The likelihood is high that a PSTN call into a branch office will be transferred from the branch AA to a local extension in the same office.
- Management philosophy allows remote locations to select their own voicemail and AA technology.

The following characteristics and guidelines apply to Cisco Unity Express in either a centralized or distributed Unified CM deployment:

- A single Cisco Unity Express can be integrated with a single Unified CM cluster.
- Cisco Unity Express integrates with Unified CM using a JTAPI application and Computer Telephony Integration (CTI) Quick Buffer Encoding (QBE) protocol. CTI ports and CTI route points control the Cisco Unity Express voicemail and automated attendant (AA) applications.
- The following CTI route points are defined on Unified CM for Cisco Unity Express:
  - Automated attendant entry point (Cisco Unity Express can contain up to five distinct AAs and may therefore require up to five different route points.)
  - Voicemail pilot number
  - Greeting management system (GMS) pilot number (Optional; if the GMS is not used, then this route point need not be defined.)
- The number of CTI ports and mailboxes supported for Cisco Unity Express on Unified CM depends on the hardware platform. For details, refer to the Cisco Unity Express data sheet available at:

  https://www.cisco.com/c/en/us/products/unified-communications/unity-express/datasheet-listing.html
- For Cisco Unity Express deployments that require more than the maximum number of supported mailboxes, consider using Cisco Unity Connection.
- Each Cisco Unity Express mailbox can be associated with a maximum of two different extensions, if needed.
- The automated attendant function for any office deployed with Cisco Unity Express can be local to the office (using the AA application in Cisco Unity Express) or centralized (using Cisco Unity Express for voicemail only).
- Cisco Unity Express can be networked with other Cisco Unity Expresses or with Cisco Unity Connection via Voice Profile for Internet Mail (VPIM) version 2. Thus, a Cisco Unity Express subscriber can send, receive, or forward messages to or from another remote Cisco Unity Express or Cisco Unity Connection subscriber.
- Cisco Unity Express allows you to specify up to three Unified CMs for failover. If IP connectivity to all three Unified CMs is lost, Cisco Unity Express switches to Survivable Remote Site Telephony (SRST) call signaling, thus providing AA call answering service as well as mailbox access to IP phones and PSTN calls coming into the branch office.
- Cisco Unity Express automated attendant supports dial-by-extension and dial-by-name functions. The dial-by-extension operation enables a caller to transfer a call to any user endpoint in the network. The dial-by-name operation uses the directory database internal to Cisco Unity Express and does not interact with external LDAP or Active Directory databases.
- Centralized Cisco Unity Express with Unified CM is not supported.
- Cisco Unity Express is not supported in pure SIP networks that do not have either Cisco Unified CM or Unified CME controlling the SIP phones.
- Cisco Unity Express can be deployed on a separate Unified CME or SRST router or a separate PSTN gateway.
- When Cisco Unity Express is deployed on a router separate from Unified CME or SRST, configure the command **allow-connections h323 to sip** for H.323-to-SIP routing.

Figure 19-14 shows the protocols involved in the call flow between Unified CM and Cisco Unity Express.

Figure 19-14 Protocols Used Between Cisco Unity Express and Unified CM

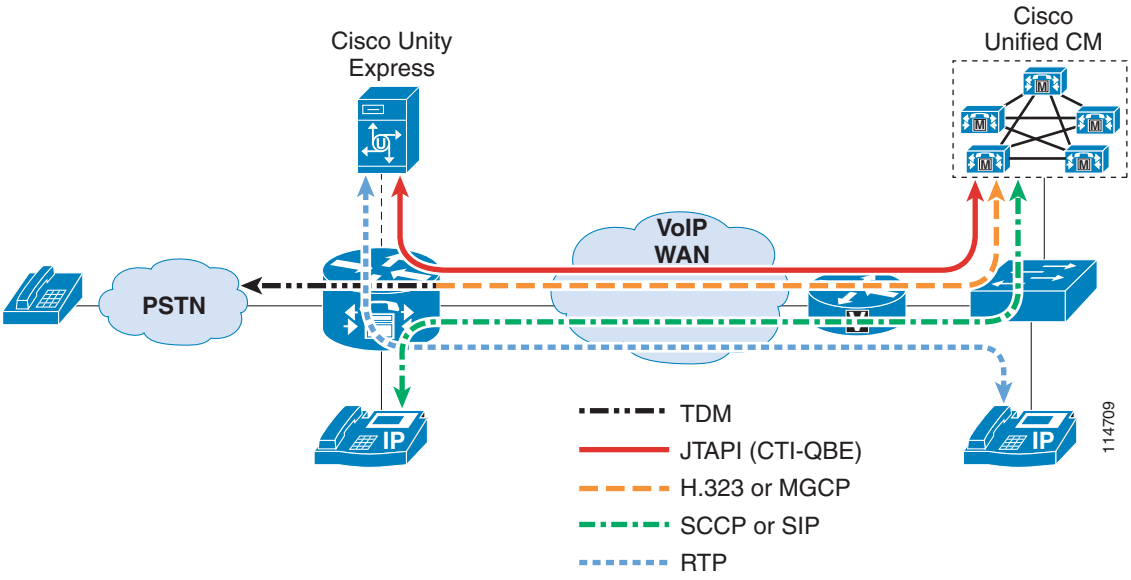


Figure 19-14 illustrates the following signaling and media flows:

- Phones are controlled via SCCP or SIP from Unified CM.
- Cisco Unity Express is controlled via JTAPI (CTI-QBE) from Unified CM.
- The Message Waiting Indicator (MWI) on the phone is affected by Cisco Unity Express communicating a change of mailbox content to Unified CM via CTI-QBE, and by Unified CM in turn sending a MWI message to the phone to change the state of the lamp.
- The voice gateway communicates via H.323, SIP, or MGCP to Unified CM.
- Real-Time Transport Protocol (RTP) stream flows carry the voice traffic between endpoints.

Figure 19-15 shows the protocols involved in the call flow between the router for SRST or E-SRST mode and Cisco Unity Express when the WAN link is down.

Figure 19-15 Protocols Used Between Cisco Unity Express and the Router for SRST or E-SRST

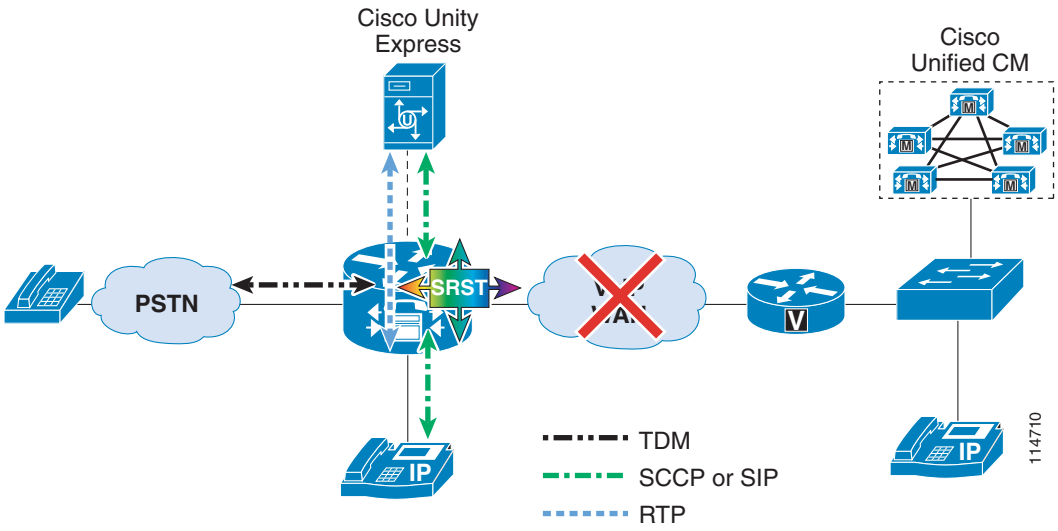


Figure 19-15 illustrates the following signaling and media flows:

- Phones are controlled via SCCP or SIP from the router for SRST or E-SRST mode.
- Cisco Unity Express communicates with the SRST router via an internal SIP interface.
- Although MWI changes are not supported in SRST mode with previous releases of Cisco Unity Express, voice messages can be sent and retrieved as during normal operation, but the MWI lamp state on the phone remains unchanged until the phone registers again with Unified CM. At that time, all MWI lamp states are automatically resynchronized with the current state of the users' Cisco Unity Express voicemail boxes. Cisco Unity Express also supports MWI for SRST mode.
- Cisco Unity Express supports SIP Subscriber/Notify and Unsolicited Notify to generate MWI notifications, in both Unified CME and SRST modes.
- RTP stream flows carry the voice traffic between endpoints.
- SRST subscribes to Cisco Unity Express for MWI for each of the ephone-dns registered to receive MWI notifications.

Note Unified CM MWI (JTAPI) is independent of the SIP MWI methods.

Voicemail Networking

This section covers specific considerations for voicemail networking, including Cisco Unity Connection and Cisco Unity Express.

Voicemail networking is the ability to allow subscribers (voicemail users) to send, receive, reply to, and forward voicemail messages between systems such as Cisco Unity Connection and Cisco Unity Express using an embedded Simple Mail Transfer Protocol (SMTP) server and a subset of the Voice Profile for Internet Mail (VPIM) version 2 protocol. Both voicemail messaging products support interoperability between one another using VPIM messaging.

Cisco Unity Express Voicemail Networking

Cisco Unity Express communicates with Cisco Unity Connection by means of VPIM for message routing and SMTP for message delivery. Cisco Unity Express voicemail networking provides the following capabilities:

- Subscribers can receive, send, and forward messages to or from another remote Cisco Unity Express or Cisco Unity Connection for locations configured on the originating system.
- Subscribers can also reply to a remote message received from a remote system.
- Subscribers can be recipients of a distribution list or individual message originating from Cisco Unity Connection.

For more information on voicemail networking with a specific product, refer to the corresponding voicemail product documentation available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-maintenance-guides-list.html

Interoperability Between Multiple Cisco Unity Connection Clusters or Networks

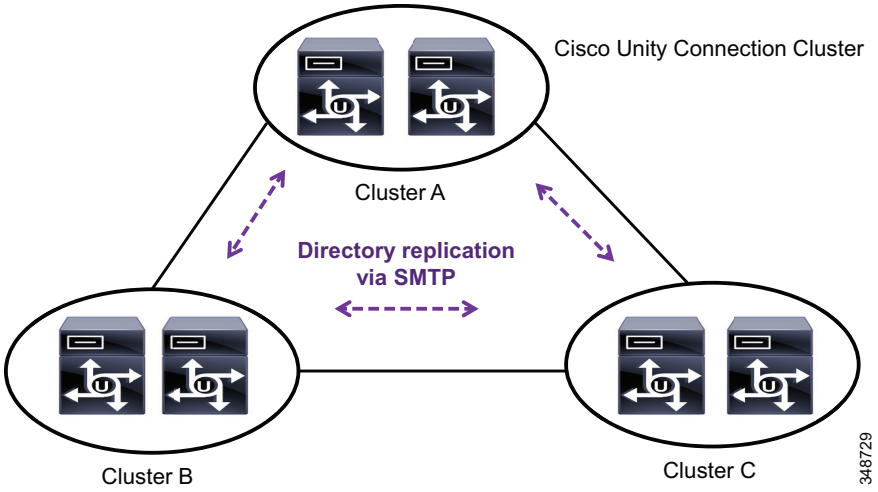
Cisco Unity Connection (digital network, HTTPS network, standalone servers, or cluster) can interoperate with another Cisco Unity Connection (digital network or HTTPS network), thus enabling users to achieve directory sharing, easy administration, and other features, as well as expanding the total number of nodes (cluster or standalone server) up to 25.

Digital Networking

Digitally networked systems use Simple Mail Transfer Protocol (SMTP) for both directory replication and message transport. As shown in Figure 19-16, multiple Unity Connection nodes are joined in full-mesh topology for sharing directory information. Only full-mesh topology is supported with Cisco Unity Connection digital networking.

Using full-mesh topology for networking requires only a single hop for transport of information between nodes, but the number of links increases with the number of nodes.

Figure 19-16 Digital Networking



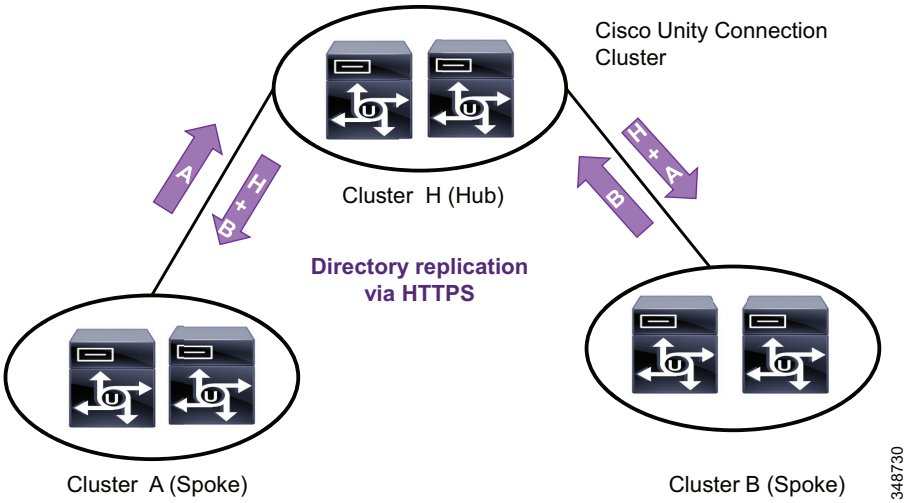
Consider the following guidelines when deploying Cisco Unity Connection digital networking:

- Each Cisco Unity Connection digital network supports a maximum of 10 servers.
- A single Cisco Unity Connection digital network supports a maximum of 100,000 users, but multiple digital networks can be joined using Voice Profile for Internet Mail (VPIM) networking to support more users. If any of the Cisco Unity Connection nodes in the digital network system is running Cisco Unity Connection 7.0, then the maximum number of users supported is 50,000.
- One Cisco Unity Connection can be a member of only one Cisco Unity Connection digital network.
- Multiple Cisco Unity Connection digital networks can be joined using VPIM. Each Cisco Unity Connection digital network must have one server defined as the bridgehead or site gateway. The bridgehead or site gateway is used to communicate with other digital networks.

HTTPS Networking

HTTPS networking uses hub and spoke topology, which enables data replication in a tree structure. The hub is a single point of communication for all the leaf spokes. All the directory replication occurs through the hub using HTTPS protocol. Each spoke is a leaf node that gathers directory information from the hub, and single spoke can connect to only one hub. As shown in Figure 19-17, spoke clusters A and B are connected to hub cluster H. If cluster A needs to fetch any directory information, it sends a query to node H. Hub node H replicates its own as well as node B's directory information to node A.

Figure 19-17 HTTPS Networking



Consider the following guidelines when deploying Cisco Unity Connection HTTPS networking:

- A single Unity Connection node or cluster can be member of only one HTTP(S) network.
- A single HTTPS network supports a maximum of 100,000 users and 150,000 contacts, but multiple digital or HTTPS networks can be joined together using Voice Profile for Internet Mail (VPIM) networking to support more than 100,000 users and/or contacts.
- A single HTTPS network system supports a single site, and each site can have a maximum of 25 nodes; however, multiple HTTPS network systems can be joined using VPIM.
- All the Cisco Unity Connection servers must be version 10.0 or higher to support HTTPS networking.
- In an HTTPS network, Cisco Unity Connection locations are joined together using a hub and spoke topology. The number of direct HTTPS links to any location must be less than or equal to 5.
- HTTPS networking cannot be used with digital networking at the same site; however, a single HTTPS network can communicate with a digital network by using VPIM. Each Cisco Unity Connection digital or HTTPS network must have one server defined as the bridgehead or site gateway. The bridgehead or site gateway is used to communicate with other digital or HTTP(S) networks.

- Full synchronization occurs after any node or cluster is added to the HTTPS network. If any discrepancy in directory data exists, then resynchronization occurs. HTTPS networking supports both manual and automatic full synchronization and resynchronization. The periodic interval for automatic synchronization is configurable.
- Directory replication occurs through the publisher node in Unity Connection. If the publisher goes down, then directory replication through this publisher node stops and a subscriber node provides directory replication.

For more information on these interoperability options, refer to the latest version of the *Networking Guide for Cisco Unity Connection*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-maintenance-guides-list.html

Cisco Unity Connection Virtualization

The Cisco Unified Computing System (UCS) is a next-generation data center platform that unites computing, networking, storage access, and virtualization into a cohesive system designed to reduce total cost of ownership (TCO) and increase business agility. Cisco Unity Connection supports virtualization over VMware with the Cisco Unified Computing system.

The following key design considerations apply to Cisco Unity Connection virtualization:

- Supports up to 20,000 users
- The Tested Reference Configurations include selected Cisco Unified Computing System (UCS) platforms. Other platforms may be supported with the specifications-based hardware support policy.
- VMware ESXi is required for virtualization.
- Servers in an active/active cluster should be on separate blades, preferably on different chassis.

Note For VMware vSphere ESXi 5.1 and earlier, at least one processor core must be reserved for the VMware ESXi hypervisor/scheduler. For VMware vSphere ESXi 5.5 and later, the Latency Sensitivity function is included to reduce virtual machine latency. When the Latency Sensitivity is set to a high value, you do not need to reserve any unused processor core for the ESXi hypervisor/scheduler.

For more information on deploying Cisco Unified Communications and Cisco Unity Connection in a virtualized system, refer to the documentation available at

http://www.cisco.com/go/virtualized-collaboration

General information about deploying Unified Communications on virtualized servers is also available in the section on Deploying Unified Communications on Virtualized Servers, page 10-55.

For Cisco Unity Connection virtualization, also refer to the latest version of the *Design Guide for Cisco Unity Connection* available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-implementation-design-guides-list.html

Best Practices for Voice Messaging

This section discusses some general best practices and guidelines that were not mentioned previously yet are important aspects of the products and should be considered in the solution. They are separated into two groupings, with Cisco Unity Connection in one grouping and Cisco Unity Express in another.

Best Practices for Deploying Cisco Unity Connection with Unified CM

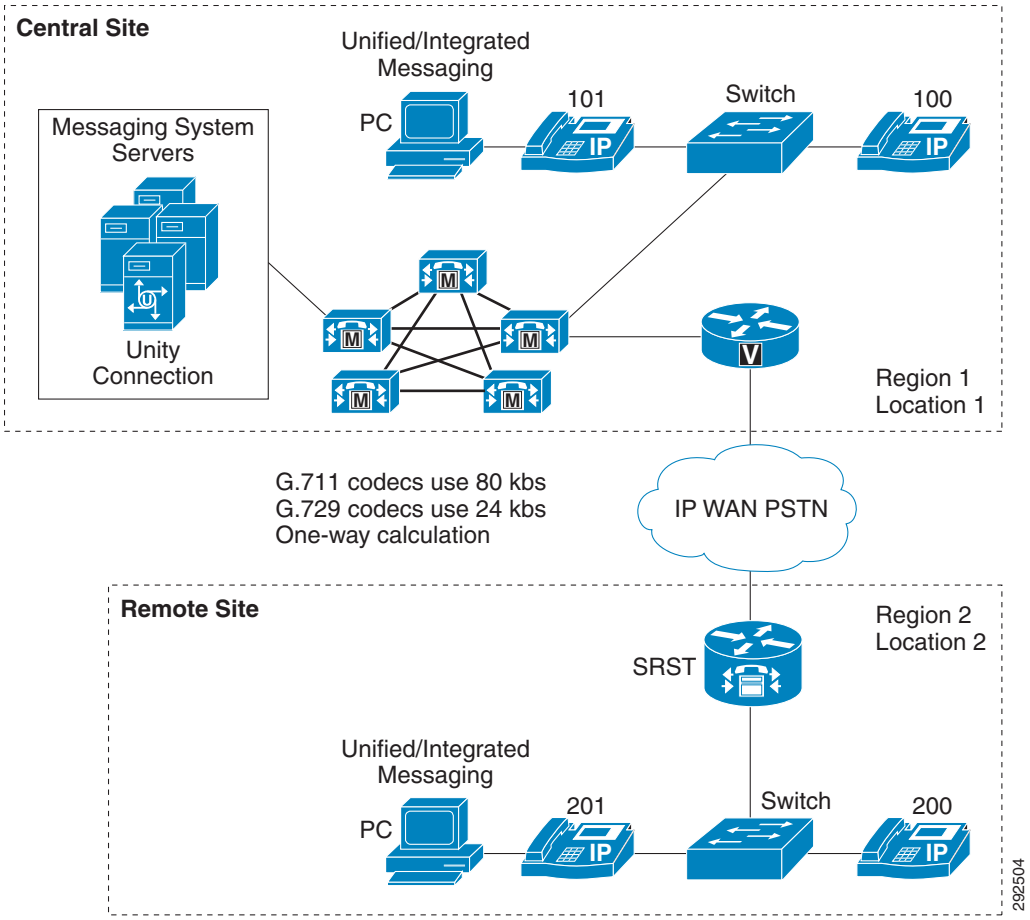
This section applies to Cisco Unity Connection. For Cisco Unity Express, see Best Practices for Deploying Cisco Unity Express, page 19-45.

Managing Bandwidth

Unified CM provides a variety of features for managing bandwidth. Through the use of regions, locations, and even gatekeepers, Unified CM can ensure that the number of voice calls going over a WAN link does not oversubscribe the existing bandwidth and cause poor voice quality. Cisco Unity Connection relies on Unified CM to manage bandwidth and to route calls. If you deploy Cisco Unity Connection in an environment where calls or voice ports might cross WAN links, these calls will be transparent to gatekeeper-based call admission control. This situation occurs any time the Cisco Unity Connection server is servicing either distributed clients (distributed messaging or distributed call processing) or when Unified CM is remotely located (distributed messaging or centralized call processing). Unified CM provides regions and locations for call admission control.

Figure 19-18 uses a small centralized messaging and centralized call processing site to illustrate how regions and locations work together to manage available bandwidth. For a more detailed discussion of regions and locations, refer to the chapter on Bandwidth Management, page 13-1.

Figure 19-18 Locations and Regions



In Figure 19-18, regions 1 and 2 are configured to use G.711 for intra-region calls and G.729 for inter-region calls. Locations 1 and 2 are both set to 24 kbps. Location bandwidth is budgeted only in the case of inter-location calls.

An intra-region (G.711) call would not be budgeted against the available bandwidth for the location. For example, when extension 100 calls extension 101, this call is not budgeted against the 24 kbps of available bandwidth for Location 1. However, an inter-region call using G.729 is budgeted against both bandwidth allocations of 24 kbps for Location 1 and Location 2. For example, when extension 100 calls extension 200, this call would be connected but any additional (simultaneous) inter-region calls would receive reorder (busy) tone.

Native Transcoding Operation

In Cisco Unity Connection, native transcoding occurs when a call is negotiated between an IP endpoint and the Cisco Unity Connection server in one codec and is recorded or played out in another codec format. If a call is negotiated in G.729 and the system-wide recording format is done in G.711, then the server has to transcode that call natively. Cisco Unity Connection native transcoding does not use external hardware transcoders but instead uses the server's main CPU. This is what is meant by native transcoding.

Cisco Unity Connection Operation

In Cisco Unity Connection, a call in any codec format supported by Cisco Unity Connection SCCP or SIP signaling (G.711 mu-law, G.711 a-law, G.729, iLBC, and G.722) will always be transcoded to Linear PCM. From Linear PCM, the recording is encoded in the system-level recording format (Linear PCM, G.711 mu-law/a-law, G.729a, or G.726), which is set system-wide in the general configuration settings (G.711 mu-law is the default). In the rest of this chapter, we refer to the codec negotiated between the calling device and Unity Connection as the "line codec," and we refer to the codec set in the system-level recording format as the "recording codec."

Because transcoding is inherent in every connection, there is little difference in system impact when the line codec differs from the recording codec. The exception to this is when using iLBC or G.722. G.722 and iLBC require more computation to transcode, therefore they have a higher system impact. G.722 and iLBC use approximately twice the amount of resources as G.711 mu-law. The subsequent impact this has is that a system can support only half as many G.722 or iLBC connections as it can G.711 mu-law connections.

As a general rule, Cisco recommends leaving the default codec as G.711. If the configuration is constrained by disk space, then a lower bit rate codec such as G.729a or G.726 can be configured as the recording format; however, keep in mind that the audio quality will not have the fidelity of G.711 audio. Also, if G.722 is used by devices on the line, then linear pulse code modulation (PCM) is an option to improve the audio quality of the recording. This will, however, increase the disk usage and impact disk space.

There are also a few reasons to change the recording codec or to choose to advertise only specific line codecs. Consider the following factors when deciding on the system-level recording format and the advertised codecs on the SCCP or SIP integration:

- Which codecs will be negotiated between the majority of the endpoints and Cisco Unity Connection? This will help you decide on which codecs need to be advertised by Cisco Unity Connection and which do not. You can then decide on when you need Unified CM to provide hardware transcoding resources in lieu of doing computationally significant native transcoding in Cisco Unity Connection, such as when requiring a large number of clients connected to Cisco Unity Connection using G.722 or iLBC.
- Which types of graphical user interface (GUI) clients (web browsers, email clients, media players, and so forth) will be fetching the recordings, and which codecs do the GUI clients support?
- What quality of the sound is produced by the selected codec? Some codecs are higher quality than others. For example, G.711 has a higher quality than G.729a, and it is a better choice if higher audio quality is necessary.
- How much disk space does the codec use per second of recording time?

Table 19-3 summarizes the characteristics of the codec formats supported by Cisco Unity Connection.

Table 19-3 Codec Characteristics

Recording Format (Codec)	Audio Quality	Supportability	Disk Space Used
Linear PCM	Highest	Widely supported	16 KBps
G.711 mu-law and a-law	Moderate	Widely supported	8 KBps
G.729a	Lowest	Poorly supported	1 KBps
G.726	Moderate	Moderately supported	3 KBps
GSM 6.10	Moderate	Moderately supported	1.6 KBps

Refer to the *System Administration Guide for Cisco Unity Connection* for details on changing the codec advertised by Cisco Unity Connection. The choices for advertised codecs are G.711 mu-law, G.711 a-law, G.729, iLBC and G.722. There is also a list of preferences according to how they are ordered in the list (top-down). For SCCP integrations, the order of the codecs has no bearing because codecs are advertised and Unified CM negotiates the codec based on the location of the port and device in the negotiated call. For SIP integrations, however, the order list is significant. If the codec is preferred, then Cisco Unity Connection will advertise that it supports both protocols but will prefer to use the one specified over the other.

For information on how to change the system-level recording format in Cisco Unity Connection Administration, refer to the *System Administration Guide for Cisco Unity Connection*.

Integration with Cisco Unified CM

Cisco Unified CM can integrate with Cisco Unity Connection via SCCP or SIP. This section discusses some specifics of that integration regarding phones, SIP trunks, and voice ports.

In Cisco Unity Connection, users are associated to a phone system that contains one or more port groups. The port groups are associated with MWI ports; thus, the MWI requests are made through the ports associated to that specific port group. Cisco Unity Connection phone systems and port groups are configured with the System Administrator.

Cisco Unity Connection supports a maximum of 90 simultaneous phone systems and port groups.Cisco recommends using a maximum of 90 port groups if you are using only the touchtone conversation (telephone user interface, or TUI) and voice recognition (voice user interface, or VUI) features of Unity Connection. If you are using all other features such as calendaring and text-to-speech (TTS), then Unity Connection supports a maximum of 60 simultaneous phone systems. These features function the same way for both SCCP and SIP integrations. For details, refer to the appropriate Cisco Unity Connection administration guides available at

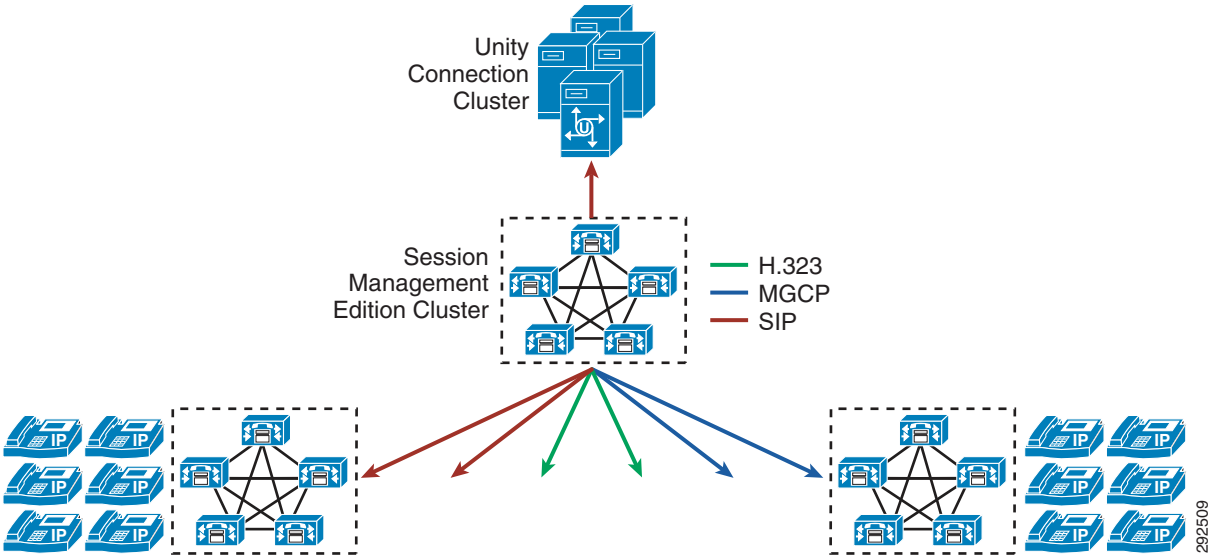
https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-maintenance-guides-list.html

In addition to the option of adding multiple clusters by adding additional integrations for each new Unified CM cluster in Cisco Unity Connection, Unified CM supports Annex M.1, Message Tunneling for QSIG, which gives administrators the ability to enable QSIG on intercluster trunks (ICTs) between Unified CM clusters. When QSIG is enabled on ICTs, Cisco Unity Connection needs to integrate with only one Unified CM cluster and designate ports only in this one cluster for turning MWIs on and off, even when supporting multiple clusters. The Annex M.1 feature in Unified CM allows for propagation of the MWI requests across the ICTs to the proper Unified CM cluster and phone within that cluster. All calls originating in other clusters can be forwarded to the Cisco Unity Connection server integrated to that one cluster. There is no need to designate MWI ports on the other clusters when Annex M.1 is enabled on the ICT.

Integration with Cisco Unified CM Session Management Edition

Cisco Unity Connection can be integrated with Cisco Unified CM Session Management Edition to provide voice messaging services to the users associated with all leaf Unified Communications clusters. (See Figure 19-19.)

Figure 19-19 Cisco Unity Connection Deployment with Unified CM Session Management Edition



The following information must be sent on the intercluster trunks between Unified Communications leaf clusters and Unified CM Session Management Edition, and on the SIP trunk to Cisco Unity Connection:

- Original called party number or redirecting number
- Calling party number
- Reason for call forward

Non- Q.SIG Trunk

For a non-Q.SIG trunk, the following settings should be enabled to deliver the original called party number or redirecting number:

- Inbound and outbound redirecting number information element (IE) delivery on MGCP and H.323 gateways and H.323 trunks
- Inbound and outbound redirecting diversion header delivery on SIP trunks

Diversion information that is sent on non-Q.SIG MGCP, H.323, or SIP trunks picks up only the calling party transformations that are defined by the voice mailbox mask of the voicemail profile that is assigned to the redirecting DN. Any calling party transformations that are defined in a route pattern or route list, or through outbound calling party transformation calling search spaces (CSSs), are not applied to diversion information.

Q.SIG-Enabled Trunk

For Q.SIG-enabled SIP, MGCP and H.323 trunks, the original called party number is sent in Q.SIG diverting leg information application protocol data units (APDUs).

On Q.SIG-enabled H.323, MGCP, and SIP trunks all calling, called, and redirecting number information is always sent in the encapsulated Q.SIG message and not in the outer H.323 message or SIP headers. The sent diversion information does not pick up any calling party transformation and does not honor any voicemail mask setting. Q.SIG tunneling-enabled trunks do not support transport of the "+" character in Q.SIG APDUs. Because of this limitation, the user's voice mailbox number should be of the same format as the directory number used in the leaf Unified Communications system. For example:

- Users with directory numbers of the format 4YYYY should have a corresponding voice mailbox number of the same 4YYYY format.
- Users with directory numbers of the E.164 format +XX4YYY should have a corresponding voice mailbox number of the same E.164 +XX4YYYY format.

Cisco Unity Connection allows an alternate extension to be associated with the voice mailbox of the user. For example:

- Primary VM box number: 4YYYY
- Alternate VM box number in +E.164: +XX4YYY

Redirected Dialed Number Information Service (RDNIS) is not supported with Q.SIG-enabled H.323 or SIP trunks. The original called party or redirecting number is sent in a Q.SIG DivertingLegInformation2 APDU instead of via RDNIS.

E.164 Number Support with Cisco Unity Connection

Cisco Unity Connection supports the E.164 number format for the following fields:

- End users' primary extensions
- Transfer rule extensions for the end users
- System call handler extensions
- Directory handler extensions
- Interview handler extensions
- Notification device phone numbers for the end users
- Personal contact phone numbers for the end users
- System contact phone numbers for the Cisco Unity Connection System
- Personal call transfer rule (PCTR) phone numbers for the Cisco Unity Connection System
- Alternate extensions for the end users
- Restriction patterns for the Cisco Unity Connection System
- Message waiting indicator (MWI) extensions for the Cisco Unity Connection System

When importing users from LDAP with E.164-formatted primary phone numbers, use the regular expression and replacement pattern that together convert phone numbers into extensions. For more information on this, refer to the sections on converting phone numbers into extensions in the latest version of the *System Administration Guide* for Cisco Unity Connection, available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-maintenance-guides-list.html

If you want to import users from Cisco Unified Communications Manager (Unified CM) with E.164 formatted extensions through AXL integration, you will have to export the E.164 extensions from Unified CM into a comma-separated values (CSV) file and perform the necessary translations on the alternate extensions (in Excel, for example) prior to using the Bulk Administration Tool (BAT) to import them into Unity Connection.

SIP URI Dialing Support with Cisco Unity Connection

Cisco Unity Connection supports SIP URI dialing for alternate extensions. SIP URI dialing enables users to access their voicemail automatically from SIP URI phones while calling to Unity Connection. A commonly used scheme for alphanumeric addresses is simplified SIP URIs of the form *user@host*, where the left-hand side (LHS, user portion) can be alphanumeric and the right-hand side (RHS, host portion) is a domain name. SIP URI is configured as an alternate extension for the user in Unity Connection.

In HTTPS and digital networking, SIP URIs for alternate extensions are replicated only at the nodes that support SIP URI.

Cisco Unity Connection SIP URI supports the following features:

- Attempt sign-in
- Ring-no-answer (RNA)
- Voicemail notification devices (work, mobile, and home phone)

An administrator can import URIs into Unity Connection from an LDAP directory or AXL integration with Cisco Unified CM.

Note The administrator cannot delete or edit an alternate extension with the Directory URI phone type. This alternate extension can be edited or deleted only from its original source (LDAP directory or Cisco Unified Communications Manager from which the user was imported).

Enhanced Message Waiting Indicator (eMWI)

Enhanced Message Waiting Indicator (eMWI) is an enhancement to traditional MWI, and it provides a visual indication of the number of voice messages. Traditional MWI works in a binary format by either enabling or disabling the message lamp on the phone whenever a new voice message arrives in or is deleted from a user's voicemail box. EMWI works with Cisco Unity Connection and is supported on the Cisco Unified IP Phones 8900 and 9900 Series SIP phones.

eMWI is a visual indication of unplayed messages in the user's voicemail box, with a colored indication depicting the status of the message. An unplayed message displays a red indication on the screen of the phone. eMWI is supported on Unified CM for Cisco Unity Connection through SIP and SCCP integrations. eMWI does not function when the system is running in SRST mode. In an integration with Cisco Unity Connection, only the messages stored on the Cisco Unity Connection servers will be indicated with eMWI, and any messages stored on an external IMAP server will not be indicated.

eMWI works in distributed call processing environments with Unified CM. In a system with distributed call processing and centralized voice messaging integration, where one cluster provides the connectivity to the voice messaging server through an intercluster trunk (H.323 or SIP), eMWI updates over the intercluster trunk are supported and are displayed on the end device. (See Figure 19-20.)

Note eMWI also works in a distributed call processing environment with centralized messaging over an intercluster trunk (H.323 or SIP).

Figure 19-20 Enhanced Message Waiting Indicator (eMWI)

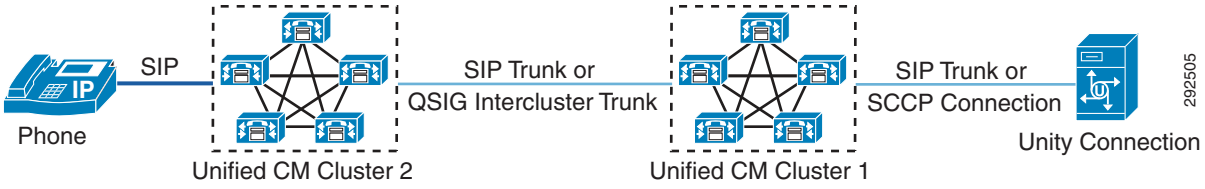
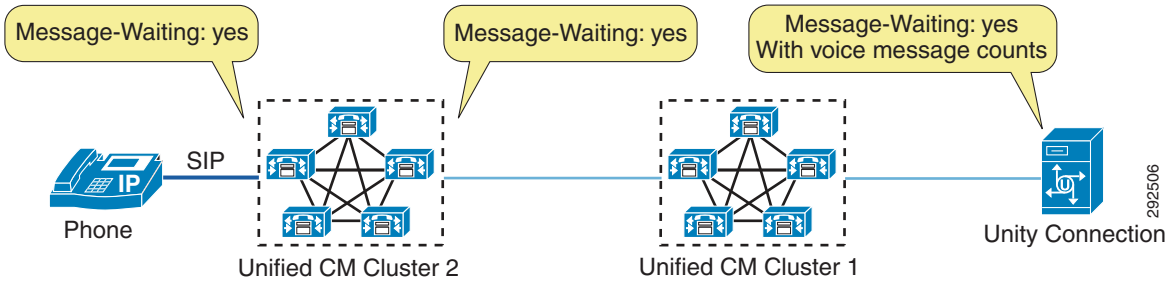


Figure 19-21 illustrates eMWI over an intercluster trunk (H.323 or SIP) in a distributed call processing environment with centralized voice messaging.

Figure 19-21 eMWI with Distributed Call Processing and Centralized Voice Messaging



As shown in Figure 19-21, Cluster 2 and its voice messaging solution support eMWI, but Cluster 1 does not. If an eMWI update with a voice message count is sent from the voice messaging solution intended for the Cluster 2 phone, Cluster 1 will forward only a standard MWI to Cluster 2 without the voice message count.

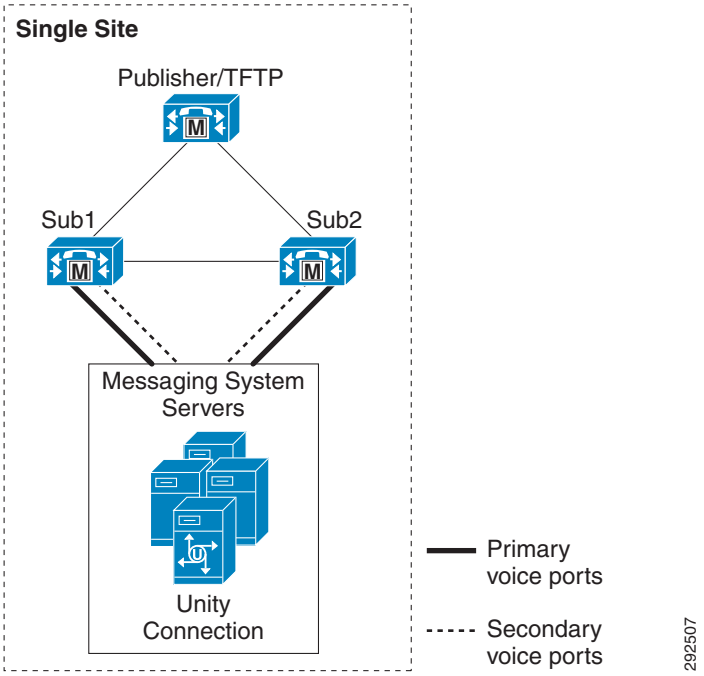
The following guidelines apply to eMWI:

- All clusters should support eMWI. If an intermediate cluster does not support eMWI, then the terminating cluster will receive a standard MWI only without voicemail counts.
- Standard MWI does not generate much traffic because it sends only a change of lamp state (ON or OFF). However, enabling eMWI can increase the amount of traffic because it also sends message counts from the messaging system. The amount of traffic depends on the number of messages and change notifications.

Voice Port Integration with a Unified CM Cluster

When deploying Cisco Unity Connection in a single-site messaging environment, integration with the Unified CM cluster occurs through the SCCP voice ports or SIP trunks. Design considerations must include proper deployment of the voice ports among the Unified CM subscribers so that, in the event of a subscriber failure (Unified CM failover), users and outside calls can continue to access voice messaging. (See Figure 19-22.)

Figure 19-22 Cisco Unity Connection Server(s) Integrated with a Unified CM Cluster (No Dedicated Backup Servers)



The Unified CM cluster in Figure 19-22 employs 1:1 server redundancy and 50/50 load balancing. During normal operations, each subscriber server is active and handles up to 50% of the total server call processing load. In the event of a subscriber server failure, the remaining subscriber server takes up the load of the failed server.

This configuration uses two groups of voicemail ports, with each group containing one-half of the total number of licensed voice ports. One group is configured so that its primary server is Sub1 and its secondary (backup) server is Sub2. The second group is configured so that Sub2 is the primary server and Sub1 is the backup.

Make sure that MWI-only ports or any other special ports are equally distributed between the two groups. During the configuration of the voice ports, pay special attention to the naming convention. When configuring the two groups of ports in Cisco Unity Connection, make sure that the device name prefix is unique for each group and that you use the same device name when configuring the voicemail ports in Unified CM Administration. The device name prefix is unique for each group of ports in this example, with group Sub1 using CiscoUM1 as the device name prefix and Sub2 using CiscoUM2 in this example.

For additional design information on the ratio of inbound to outbound voicemail ports (for MWI, message notification, and TRaP), refer to the latest version of the Cisco Unity Connection *System Administration Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-maintenance-guides-list.html

Note The device name prefix is unique for each group of ports and must match the same naming convention for the voicemail ports configured in Unified CM Administration.

In Unified CM Administration, half of the ports in this example are configured to register using the unique device name prefix of CiscoUM1, and the other half are configured to register using the unique device prefix CiscoUM2. (See Table 19-4.) When the ports register with Unified CM, half will be registered with subscriber server Sub1, and the other half will be registered with Sub2, as shown in Table 19-4.

Table 19-4 Voicemail Port Configuration in Unified CM Administration

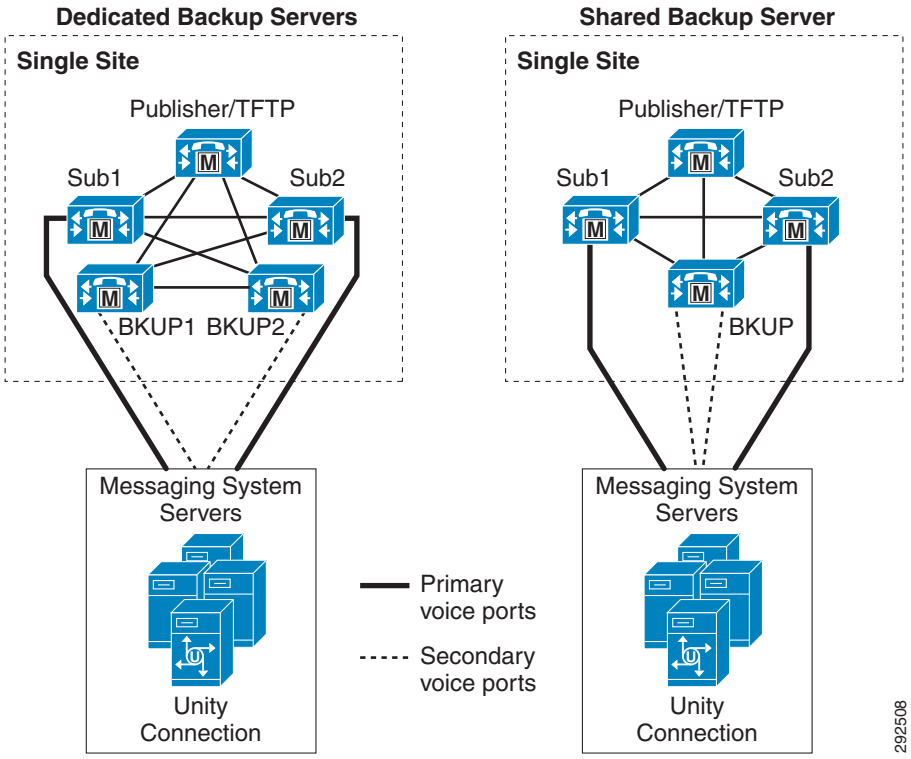
Device Name	Description	Device Pool	SCCP Security Profile	Status	IP Address
CiscoUM1-VI1	Unity Connection 1	Default	Standard Profile	Registered with sub1	1.1.2.9
CiscoUM1-VI2	Unity Connection 1	Default	Standard Profile	Registered with sub1	1.1.2.9
CiscoUM1-VI3	Unity Connection 1	Default	Standard Profile	Registered with sub1	1.1.2.9
CiscoUM1-VI4	Unity Connection 1	Default	Standard Profile	Registered with sub1	1.1.2.9
CiscoUM2-VI1	Unity Connection 1	Default	Standard Profile	Registered with sub2	1.1.2.9
CiscoUM2-VI2	Unity Connection 1	Default	Standard Profile	Registered with sub2	1.1.2.9
CiscoUM2-VI3	Unity Connection 1	Default	Standard Profile	Registered with sub2	1.1.2.9
CiscoUM2-VI4	Unity Connection 1	Default	Standard Profile	Registered with sub2	1.1.2.9

Note The naming convention used for the voicemail ports in Unified CM Administration must match the device name prefix used in Cisco UTIM, otherwise the ports will fail to register.

Voice Port Integration with Dedicated Unified CM Backup Servers

This Unified CM cluster configuration allows each subscriber server to operate at a call processing load higher than 50%. Each primary subscriber server has either a dedicated or shared backup server. (See Figure 19-23.) During normal operation, the backup server processes no calls; but in the event of failure or maintenance of a Subscriber server, the backup server will then take the full load of that server.

Figure 19-23 Cisco Unity Connection Server(s) Integrated with a Single Unified CM Cluster with Backup Subscriber Server(s)



Configuration of the voicemail ports in this case is similar to the 50/50 load-balanced cluster. However, instead of configuring the voice ports to use the opposite subscriber server as the secondary server, the individual shared or dedicated backup server is used. In the Unified CM cluster with a shared backup server, both of the secondary ports for the subscriber servers are configured to use the single backup server.

The voice port names (device name prefix) must be unique for each Cisco UTIM group and must be the same as the device names used on the Unified CM server.

To configure the voicemail ports on Cisco Unity Connection, use the Telephony Integration section of the Unity Connection Administration console. For details, refer to the Cisco Unity Connection administration guides available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-maintenance-guides-list.html

IPv6 Support with Cisco Unity Connection

The current requirements for IP addressing are surpassing the available set of IP address with IPv4, the current version of IP addressing. Therefore, most IP-based solutions are moving toward incorporating support for IPv6, which provides many more available IP addresses than IPv4. Cisco Unity Connection supports IPv6 addressing with Cisco Unified Communications Manager system integrations through SCCP or SIP. At a component level, dual-stack addressing (both IPv4 and IPv6) is supported over call control and media only.

Cisco Unity Connection supports following IPv6 address types:

- Unique Local Address
- Global Address

Note Voice messages are stored as .wav files and are independent of IPv6 or IPv4.

IPv6 support is disabled by default, but system administrators can enable IPv6 and configure IPv6 address settings either in Cisco Unified Operating System Administration or in the command line interface (CLI). Cisco Unity Connection can obtain an IPv6 address either through router advertisement, through DHCP, or from addresses configured manually either in Cisco Unified Operating System Administration or through the CLI. Cisco Unity Connection Administration and Cisco Personal Communications Assistant can be accessed using IPv6 addresses.

Note IPv6 addressing cannot be enabled during installation or upgrade of Cisco Unity Connection. Cisco Unity Connection does not support "IPv6 ONLY" server configuration. Cisco Unity Connection supports Unicast only for IPv6.

Cisco Unity Connection over IPv6 supports following functionality:

- Cisco Unity Connection offers auto-discovery functionality over IPv6, which allows Unity Connection to search for Microsoft Exchange servers to communicate with.
- Cisco Unity Connection can be integrated with an IPv6 Microsoft Exchange 2007 or 2010 server to enable the Single Inbox feature.
- Cisco ViewMail for Outlook (VMO) supports communication between Outlook and Cisco Unity Connection over IPv6.
- Voice messages received on Cisco Unity Connection can be accessed using any IMAP client such as Outlook over IPv6.
- Cisco Unity Connection can be integrated with LDAP over IPv6 to import the user information.
- Cisco Unity Connection also offers Telephone Record and Playback (TRaP) functionality over IPv6, which enables users to record or play back messages over an IPv6-enabled phone so that signaling can happen over IPv6.

Single Inbox with Cisco Unity Connection

Cisco Unity Connection supports the Single Inbox feature with Microsoft Exchange 2003, 2007, and 2010 (clustered or non-clustered), thereby providing Unified Messaging for voicemail. Cisco Unity Connection can support all three of these Microsoft Exchange versions simultaneously or any one of them separately. Unity Connection also supports interoperability with the Microsoft Business Productivity Online Suite (BPOS)-Dedicated Services and Microsoft Office 365 cloud-based exchange server. Unity Connection uses Microsoft Exchange Online to enable the Single Inbox feature. For more information, refer to the latest version of the *Unified Messaging Guide for Cisco Unity Connection*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-maintenance-guides-list.html

All voice messages, including those sent from Cisco Unity Connection ViewMail for Microsoft Outlook, are first stored in Cisco Unity Connection and are immediately replicated to the Microsoft Exchange mailbox for the recipient; however, replication is optional. Also, this feature can be configured per individual user.

Cisco Unity Connection support of Unified Messaging for voicemail involves several design considerations. The user's email becomes a single container for all messages, including email and voicemail. If a message is moved to any other folder under the user's Inbox, it will continue to show up in Cisco Unity Connection. However, if the user moves voice messages into Outlook folders that are not under the Inbox folder, the messages are deleted from Cisco Unity Connection but they can still be played by using ViewMail for Outlook because a copy still exists in the Outlook folder. If the user moves the messages back into the Inbox folder or into a folder under the Inbox folder, the message is synchronized back into the Cisco Unity Connection mailbox for that user. In addition, when a user deletes a voice message from Cisco Unity Connection or when Cisco Unity Connection automatically deletes a voice message because of message aging, the message is also deleted from Microsoft Exchange. Likewise, when a voice message is deleted from Microsoft Exchange, it is also deleted from Cisco Unity Connection.

If a message is marked as secured and private, the actual message is not replicated in Microsoft Exchange; instead, a placeholder with a brief description is created for the message. The only copy of actual message stays on Cisco Unity Connection, and when the user retrieves the message it is played back from Cisco Unity Connection directly instead of from a local source, unlike in the case of a normal message. This also means that there is no local access to the audio file if it is accessed through voicemail from Outlook. Movement of the secure and private message to any folder other than Inbox and folders below Inbox would result in deletion of the message permanently, thereby leaving no opportunity for retrieval.

Note All voice messages remain on the Cisco Unity Connection server regardless of the type of messaging deployment. Cisco Unity Connection is the authoritative source of voice messaging traffic, notifications, and synchronizations.

The amount of space a single voicemail message can acquire is configured on the Cisco Unity Connection server and is similar to message aging. The maximum size for a voicemail message is also configured on the Microsoft Exchange Server. Typically, the Microsoft Exchange Server maintains a larger size than Cisco Unity Connection that is synchronized to the mailbox. Hence, the minimum size of the message in Microsoft Exchange should be bigger than the maximum size in Cisco Unity Connection.

From a security standpoint for communications between Cisco Unity Connection and Microsoft Exchange, HTTPS is chosen as the default option. HTTP is also supported but not recommended because it reduces security and might also need further configuration on Microsoft Exchange. At the same time, there is an option to validate the Microsoft Exchange certificate, provided that access to the certificate server is available.

Best Practices for Deploying Cisco Unity Express

When deploying Cisco Unity Express, use the following guidelines and best practices:

- Ensure that the IP phones having Cisco Unity Express as their voicemail destination are located on the same LAN segment as the router hosting Cisco Unity Express.
- If uninterrupted automated attendant (AA) and voicemail access is required for a site deployed with Cisco Unity Express, ensure that Cisco Unity Express, SRST, and the PSTN voice gateway are all located at the same physical site. Hot Standby Router Protocol (HSRP) or other redundant router configurations are not currently supported with Cisco Unity Express.
- Each mailbox can be associated with a primary extension number and a primary E.164 number. Typically, this number is the direct-inward-dial (DID) number that PSTN callers use. If the primary E.164 number is configured to any other number, use Cisco IOS translation patterns to match either the primary extension number or primary E.164 number so that the correct mailbox can be reached during SRST mode.

Voicemail Integration with Unified CM

- Each Cisco Unity Express site must be associated with a CTI route point for voicemail and one for AA (if licensed and purchased), and you must configure the same number of CTI ports as Cisco Unity Express ports licensed. Ensure that the number of sites with Cisco Unity Express does not exceed the CTI scalability guidelines presented in the chapter on Call Processing, page 9-1.
- Cisco Unity Express is associated with a JTAPI user on Unified CM. Although a single JTAPI user can be associated with multiple instances of Cisco Unity Express in a system, Cisco recommends associating each dedicated JTAPI user in Unified CM with a single Cisco Unity Express.
- If Unified CM is upgraded from a previous version, the password of the JTAPI user automatically gets reset on Unified CM. Therefore, after the upgrade, the administrator must make sure that the JTAPI password is synchronized between Cisco Unity Express and Unified CM so that Cisco Unity Express can register with Unified CM.
- The CTI ports and CTI route points can be defined in specific locations. Cisco recommends using locations-based call admission control between Unified CM and Cisco Unity Express. RSVP may also be used.
- Ensure proper Quality of Service (QoS) and bandwidth for signaling traffic that traverses the WAN between Cisco Unity Express and Unified CM. Provision 20 kbps of bandwidth for CTI-QBE signaling for each Cisco Unity Express site. See the chapter on Network Infrastructure, page 3-1, for more details.
- The CTI-QBE signaling packets from Unified CM to Cisco Unity Express are marked with a DSCP value of AF31 (0x68). Unified CM uses TCP port 2748 for CTI-QBE signaling.
- The Unified CM JTAPI library sets the proper IP Precedence bits in all outgoing QBE signaling packets. As a result, all signaling between Cisco Unity Express and Unified CM will have the proper QoS bits set.

Cisco Unity Express Codec and DTMF Support

Calls into Cisco Unity Express use G.711 only. Cisco recommends using a local transcoder to convert the G.729 calls traversing the WAN into G.711 calls. You can configure Unified CM regions with the G.711 voice codec for intra-region calls and the G.729 voice codec for inter-region calls.

If transcoding facilities are not available at the Cisco Unity Express site, provision enough bandwidth for the required number of G.711 voicemail calls over the WAN. Configure the Unified CM regions with the G.711 voice codec for calls between the IP phones and Cisco Unity Express devices (CTI ports and CTI route points).

Cisco Unity Express does not support in-band DTMF tones; it supports only DTMF relay. With Cisco Unity Express, DTMF is carried out-of-band via either the SIP or JTAPI call control channels. Cisco Unity Express 2.3 supports G.711 SIP calls with RFC 2833 into Cisco Unity Express.

JTAPI, SIP Trunk, and SIP Phone Support

Cisco Unified CM supports SIP trunking protocol; however, Cisco Unity Express uses JTAPI to communicate with Unified CM. Cisco Unity Express supports both SCCP and SIP phones.

- Configure a SIP trunk for SRST and Unified CM for support of SIP phones (through JTAPI).
- Cisco Unity Express supports G.729 SIP calls via a transcoder, with the ability added in Cisco IOS Release 12.3(11)XW for RFC 2833 to pass through a transcoder.
- Cisco Unity Express supports delayed media (no SDP in the INVITE message) for call setup in case of a slow-start call from Unified CM.
- Cisco Unity Express supports both blind and consultative transfer, but the default transfer mode is consultative transfer (semi-attended) using REFER in SIP calls. Use the Cisco Unity Express command line interface to explicitly change the transfer mode to consultative transfer using REFER or blind transfer using BYE/ALSO. If REFER is not supported by the remote end, BYE/ALSO will be used.
- Cisco Unity Express supports outcall for voice message notifications. It also supports consultative transfers. During both of these call setups, Cisco Unity Express can receive 3*xx* responses to the INVITE. Cisco Unity Express processes only 301 (Moved Permanently) and 302 (Moved Temporarily) responses to the INVITE. This requires the URL from the Contact header from the 3*xx* response to be used to send a new INVITE. 305 (Use Proxy) responses are not supported.

Note For compatibility between Cisco Unified CM and Cisco Unity Express, refer to the *Cisco Unity Express Compatibility Matrix*, available at https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/unity_exp/compatibility/cuecomp.html.

For more information about Cisco Unity Express, refer to the product documentation available at

https://www.cisco.com/c/en/us/products/unified-communications/unity-express/index.html

Third-Party Voicemail Design

This section discusses various options for deploying third-party voicemail systems with Cisco Unified Communications, and it covers both integration and messaging.

Note This section does not discuss how to size a third-party voicemail system for ports and/or storage. For this type of information, contact your voicemail vendor, who should be better able to discuss the individual requirements of their own system, based upon specific traffic patterns.

Integration

Integration is defined as the physical connection between a voicemail system and its associated PBX or call processing agent, and it also provides for the feature set between the two. There are many voicemail vendors, and it is not uncommon for customers to want to continue to use an existing voicemail system when deploying Cisco Unified CM.

Note Cisco does not test or certify any third-party voicemail systems. Within the industry, it is generally considered to be the responsibility of the voicemail vendor to test and/or certify their products with various PBX systems. Cisco does, of course, test its interfaces to such equipment and will support these interfaces regardless of which third-party voicemail system is connected.

Cisco Unified CM can be integrated with a third-party PBX by using QSIG, which also allows a third-party PBX to connect to Unified CM through a Primary Rate Interface (PRI) T1/E1 trunk. Each method has its own advantages and disadvantages, and the method you employ will largely depend on how your voicemail system is integrated to your current PBX.

Today there are other potential methods of voicemail integration, such as H.323 or SIP. However, due to the varying methods of vendor implementation, features supported, and other factors, these third-party voicemail integrations will have to be evaluated on a per-customer basis. Customers are advised to contact their Cisco Account Team and/or Cisco Partner to discuss these options further.

Messaging

Messaging is defined as the exchange of messages between voicemail systems, and there are several open standards for this purpose.

The most common protocol deployed to allow messaging between dissimilar systems is Voice Profile for Internet Mail (VPIM). VPIM has seen several updates to its specification, and although Version 2 is not the latest, it still appears to be the most widely adopted. The messaging protocol prior to VPIM is Audio Messaging Interchange Specification - Analog (AMIS-A), and it is fairly rare in its adoption due mainly to its cumbersome user interface as well as the analog technology it employs and its lack of features.

Collaboration Instant Messaging and Presence

Revised: March 1, 2018

Cisco Unified Communications Manager IM and Presence Service provides native standards-based, dual-protocol, enterprise instant messaging (IM) and network-based presence as part of Cisco Unified Communications. This secure, scalable, and easy-to-manage service within Cisco Unified Communications Manager offers users feature-rich communications capabilities both within and external to the enterprise.

The Cisco Unified Communications Manager IM and Presence Service is one of the options for IM and Presence, which enhances the value of a Cisco Unified Communications and Collaboration system. The main presence component of the solution is the Cisco IM and Presence Service for all on-premises deployment needs, which incorporates the Extensible Communications Platform (XCP) and supports SIP/SIMPLE and Extensible Messaging and Presence Protocol (XMPP) for collecting information regarding a user's availability status and communications capabilities. The user's availability status indicates whether or not the user is actively using a particular communications device such as a phone. The user's communications capabilities indicate the types of communications that user is capable of using, such as video conferencing, web collaboration, instant messaging, or basic audio.

The IM and Presence Service is tightly integrated with Cisco and third-party compatible desktop and mobile presence and instant messaging clients, which also include Cisco Jabber SDK. It enables the clients to perform various functions such as instant messaging, presence, click-to-call, phone control, voice, video, visual voicemail, and web collaboration. The IM and Presence Service offers the flexibility of rich, open interfaces that enable implementation for IM and Cisco network-based presence, as well as IM and presence federation for a wide variety of business applications.

The aggregated user information captured by the Cisco IM and Presence Service enables Cisco Jabber, Cisco Unified Communications Manager applications, and third-party applications to increase user productivity. These applications help connect colleagues more efficiently by determining the most effective form of communication.

Note The Cisco IM and Presence Service must be deployed with the equivalent version of Cisco Unified Communications Manager (Unified CM) using Cisco provided VM configuration OPTION user configuration templates on virtual servers on Cisco Unified Computing System (UCS). Cisco does not support over-subscription of VM resources between virtual servers, and system resources should be dedicated per virtual server being deployed.

This chapter explains the basic concepts of presence and instant messaging within the Cisco Unified Communications System for on-premises, cloud, and hybrid options, and it provides guidelines for how best to deploy the various components of the presence and instant messaging solution.

What's New in This Chapter

Table 20-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 20-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Persistent chat	Persistent Chat and Compliance Logging Considerations, page 20-31	March 1, 2018
Centralized IM and Presence deployments	Centralized IM and Presence Deployments, page 20-32	March 1, 2018

Presence

Presence refers to the ability and willingness of a user to communicate across a set of devices. It involves the following phases or activities:

- Publish user status

  User status changes can be published automatically by recognizing user keyboard activity, phone use, WebEx Meeting status, device connectivity to the network, and calendar status from Microsoft Exchange.

- Collect this status

  The published information is gathered from all the available sources, privacy policies are applied, and then current status is aggregated, synchronized, and stored for consumption.

- Consume the information

  Desktop applications, calendar applications, and devices can use the user status information to provide real-time updates for the end users to make better communication decisions.

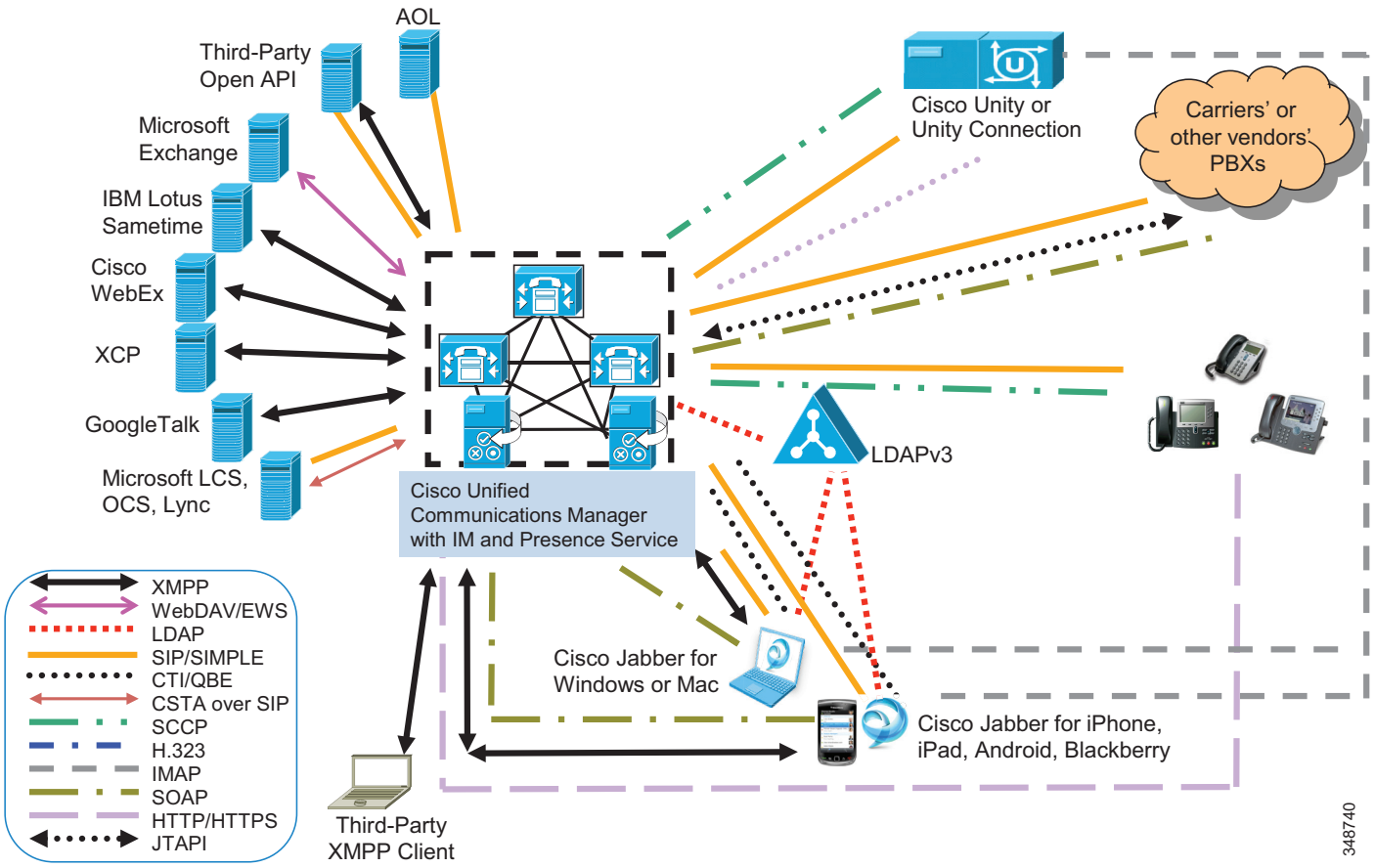
Status combines the capabilities of what the device or user can do (voice, video, instant messaging, web collaboration, and so forth) and the attributes showing the state of the device or user (available, busy, on a call, and so forth). Presence status can be derived from automatic events such as client login and telephone off-hook, or it can be derived from explicit notification events for changing status such as the user selecting Do Not Disturb from a change-status pick list.

Terminology surrounding presence refers to a watcher, presence entity (*presentity*), and presence server. The presence entity publishes its current status to the presence server by using a PUBLISH or REGISTER message for SIP/SIMPLE clients, or by using an XML Presence Stanza for XMPP clients. It can be a directory number (DN) or a SIP uniform resource identifier (URI) that resides within or outside the communications cluster. A *watcher* (device or user) requests presence status about a presence entity by sending a message to the presence server. The presence server responds to the watcher with a message containing the current status of the presence entity.

On-Premises Cisco IM and Presence Service Components

Cisco IM and Presence Service encompasses the components illustrated in Figure 20-1.

Figure 20-1 Cisco IM and Presence Service Interfaces



On-Premises Cisco IM and Presence Service User

For presence, typically a user is described in terms of the user's presence status, the number of users on the system, or the user's presence capabilities.

A user, specified in Unified CM as an *end user*, has to be associated with a line appearance. When using the IMP PUBLISH Trunk service parameter on Unified CM, you must associate the user with a line appearance. With the line appearance, the user is effectively tied to a line appearance (directory number or URI associated with a particular device), which allows for a more detailed level of granularity for aggregation of presence information. The user can be mapped to multiple line appearances, and each line appearance can have multiple users (up to 5).

Note For voice-only, video-only, or IM-only deployments, the IM and Presence Service in Cisco Collaboration System Release (CSR) 12.*x* has a cluster capacity limit of 75,000 users for Full UC Mode.

The concept of a *presence user* appears throughout this chapter; therefore, keep in mind the meaning of a user as defined for Cisco IM and Presence. By default an IM and Presence Service user is defined in a Unified Communications deployment as *user@default_domain* (the basis for the Jabber Identifier, or JID), where *user* is what is configured manually or in the Unified CM LDAP synchronization agreement (sAMAccountName, email, employeenumber, telephonenumber, or UserPrincipalName) and *default_domain* is the domain configured in the IM and Presence Service administration.

Enhanced IM Addressing and IM Address Schemes

The Enhanced IM Addressing feature provides for additional IM Address (JID) configuration options on Unified CM IM and Presence, which provides multi-domain support.

IM Addressing Schemes:

- *UserID@Default_Domain* is the default IM addressing scheme when you install the IM and Presence Service.
- DirectoryURI IM addressing scheme supports multiple domains, alignment with the user's email address, and alignment with Microsoft SIP URI.

The default setting of user@default_domain allows for only a single domain, whereas DirectoryURI allows for greater flexibility in handling multiple domains and email addresses as the contact identifier. A user can log into Jabber with their sAMAccountName attribute, while the Jabber ID is mapped to the DirectoryURI field. The Fleixble JID structure makes it independent of UID for authentication.

Note DirectoryURI is a global administrative setting on Unified CM. If DirectoryURI is selected for IM and Presence Service addressing, all clients in the deployment must be able to handle and support the DirectoryURI option.

While UserID can be mapped to the email address, that does not mean the IM URI equals the email address. Instead it becomes *<email-address>@Default_Domain*. For example, amckenzie@example.com@sales-example.com. The Active Directory (AD) mapping setting that you choose is global to all users within that IM and Presence Service cluster. It is not possible to set different mappings for individual users.

Unlike the UserID@Default_Domain IM addressing scheme, which is limited to a single IM domain, the DirectoryURI IM addressing scheme supports multiple IM domains. Any domain specified in the DirectoryURI is treated as hosted by the IM and Presence Service. The user's IM address is used to align with their DirectoryURI, as configured on Cisco Unified Communications Manager.

Single Sign-On (SSO) Solutions

Cisco recommends the use of SAML SSO (Release 10.0(1) and later) as the Single Sign-On (SSO) solutions.

SAML is an open standard that enables federated authentication across all operating systems. The SAML standard enables clients to authenticate against any SAML-enabled Collaboration service regardless of the operating system (OS) of the client's platform. SAML is a set of standards that have been defined to share information about who a user is and what the user’s attributes are, as well as providing a way to request authentication and grant or deny access to something. For example, two different organizations can use SAML to establish trust relations without exchanging passwords.

SAML supports the following end-user applications:

- Cisco Unified CM
- Cisco IM and Presence Service
- Cisco Unity Connection
- Cisco WebEx Connect and Messenger Cloud

SAML SSO also supports the following end-user clients:

- WebEx iOS
- WebEx Android
- WebEx Connect
- WebEx Messenger
- Jabber for Windows
- Jabber iOS
- Jabber for Android
- Jabber for Mac

For information about SAML SSO, see the section on On-Premises Cisco IM and Presence Service SAML SSO for Jabber, page 20-40, or the latest version of the *SAML SSO Deployment Guide for Cisco Unified Communications Applications* available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

IM and Presence Collaboration Clients

The Cisco Collaboration software family of Jabber clients allows users to easily access capabilities for voice and video calls, provides a contact directory with presence information for colleagues, and includes tools for instant messaging (IM), voice messaging, desktop sharing, and conferencing.

Cisco Jabber clients offer various options to choose from, as well as third-party XMPP clients and applications that can be used. Cisco Jabber clients integrate with underlying Cisco Unified Communication services through a common set of interfaces. In general, each client provides support for a specific operating system and both desktop and mobile applications.

The client-specific sections of this chapter also provide relevant deployment considerations, planning, and design guidance around integration into the Cisco Unified Communications System.

The following Collaboration clients are supported by the Cisco Unified Communications System:

- Desktop clients include:
  - Cisco Jabber for Windows
  - Virtualization Experience Media Engine (VXME) for Windows
  - Cisco Jabber for Mac
- Web platforms include:
  - Cisco Jabber Guest
  - Cisco Jabber Web SDK

- Mobile clients include:
  - Cisco Jabber for iPhone
  - Cisco Jabber for iPad
  - Cisco Jabber for Blackberry
  - Cisco Jabber for Android
  - Mobile and Tablet

Cisco Jabber desktop clients for both Macintosh and Windows provide robust and feature-rich collaboration capabilities, including standards-based IM and presence, audio and video, visual voicemail, desktop sharing, desk phone control, Microsoft Office integration, and contact management.

Cisco Jabber desktop clients can be deployed to use on-premises services in which Cisco IM and Presence and Cisco Unified Communications Manager provide client configuration, instant messaging and presence, and user and device management. Cisco Jabber for Windows and Cisco Jabber for Mac can also be deployed to use cloud-based services through integration with Cisco WebEx Messenger service.

Multiple Device Messaging (MDM) and Logins

All Cisco Jabber clients are associated to an IM and Presence Service node when a user logs in on any of the Jabber desktop or mobile clients. The IM and Presence node to which the user is associated monitors all changes for that user such as availability, contact list, and feature-based tasks.

The IM and Presence Service node tracks all of the registered clients for every presence-enabled user, whether they are logged in from one or all of the various Jabber clients such as Cisco Jabber for iOS, Android, Windows, or Mac.

When a new IM session is initiated between users who are logged into multiple clients, the first incoming message is broadcast to all of the registered clients of the receiving user. The IM and Presence Service node then waits for the first response from one of the registered clients. The first client to respond subsequently receives the remainder of the incoming messages until the user starts responding from another registered client. The node then reroutes subsequent messages to this new client.

For example:

> Johnny wishes to initiate an IM conversation with Betty. He has previously logged into Cisco Jabber for Windows and Cisco Jabber for Android. Betty has registered two clients with the central IM and Presence Service node. Johnny initiates the conversation by sending the message, "Hi Betty. Are you free?"
>
> The IM and Presence Service node identifies that Betty has two registered clients, and it broadcasts Johnny's message to both. Betty is sitting at her desk and observes Johnny's message appearing on both her laptop and phone. She chooses to respond using her laptop and responds with the message, "I have a meeting in a few moments but I can chat briefly right now."
>
> The IM and Presence Service node identifies that Betty has responded using Cisco Jabber for Windows, and it marks this as the client to which to route all subsequent messages in the conversation. When Johnny responds with, "This will only take a minute," this response is routed directly to Cisco Jabber for Windows. If Betty starts responding to Johnny using her phone at some point in the conversation, the IM and Presence Service node will then route subsequent messages there instead of to Cisco Jabber for Windows.

Note Every client a user logs into affects the total capacity limits for users and devices on the IM and Presence and Unified CM cluster. For example, on a cluster with a 15k-user VM configuration, if every user logs into their iPhone and desktop Jabber clients, then the maximum capacity would be 7,500 presence users instead of 15,000.

Jabber Desktop Client Modes

Cisco Jabber desktop clients can operate in one of two modes for call control:

- Softphone Mode — Using audio and video on a computer

  When a Jabber desktop client is in softphone mode, it is directly registered to Unified CM as a SIP endpoint for audio and video call control functionality, and it is configured on Unified CM as device type **Client Services Framework**.

- Deskphone Control Mode — Using a Cisco IP Phone for audio (and video, if supported)

  When a Jabber desktop client is in deskphone control mode, it does not register with Unified CM using SIP, but instead it uses CTI/JTAPI to initiate, monitor, and terminate calls, monitor line state, and provide call history, while controlling a Cisco IP Phone.

Cisco Jabber for Mobile Clients

Cisco provides collaboration clients for the following mobile devices: Android, BlackBerry, and Apple iOS devices such as iPhone and iPad. For more information on Cisco Jabber for mobile devices, see the chapter on Mobile Collaboration, page 21-1.

Cisco UC Integration™ for Microsoft Lync

Cisco UC Integration™ for Microsoft Lync allows for integrated Cisco Unified Communications services with Microsoft Lync and Microsoft Office Communications Server (OCS) R2, while delivering a consistent user experience. The solution extends the presence and instant-messaging capabilities of Microsoft Lync by providing access to a broad set of Cisco Unified Communications services, including standards-based audio and video, unified messaging, web conferencing, deskphone control, and telephony presence.

Third-Party XMPP Clients and Applications

Cisco IM and Presence, with support for SIP/SIMPLE and Extensible Messaging and Presence Protocol (XMPP), provides support of third-party clients and applications to communicate presence and instant messaging updates between multiple clients. Third-party XMPP clients allow for enhanced interoperability across various desktop operating systems. In addition, web-based applications can obtain presence updates, instant messaging, and roster updates using the HTTP interface with SOAP, REST, or BOSH (based on the Cisco AJAX XMPP Library API). For additional information on the third-party open interfaces, see the section on Third-Party Presence Server Integration, page 20-62.

SAML Single Sign On

Single Sign-On allows Cisco Jabber users to securely access all Jabber services without being prompted to log into each of them separately. The Cisco Jabber application uses authentication performed by the corporate Identity Provider. The Identity Provider can control the authentication experience for Cisco

Jabber users – for example, by prompting users for their enterprise username and password once when the Cisco Jabber application is first run and by specifying the length of time a user is authorized to use Cisco Jabber services.

The Cisco Jabber application uses the Security Assertion Markup Language (SAML), which is an XML-based open standard data format that enables access to a defined set of Cisco services transparently after verifying credentials with an Identity Provider. SAML Single Sign-On can be enabled for Cisco WebEx Messenger Services, Cisco Unified Communications Manager, and Cisco Unity Connection. SSO is deployed for use with Cisco Jabber clients using service discovery.

SAMLv2 SSO supports the following deployment models:

- On-premises deployment models
  - IM and Presence and Unified CM
  - IM and Presence, Unified CM, and Unity Connection (SSO)
  - IM and Presence, Unified CM, Unity Connection (SSO), and Cisco Mobile Workspace Solution (SSO or not)
- Hybrid deployment models
  - WebEx Messenger (SSO) and Unified CM
  - WebEx Messenger (SSO), Unified CM, and Unity Connection
  - WebEx Messenger (SSO), Unified CM, Unity Connection, and WebEx Meeting Center

Cisco Unified CM User Data Service (UDS)

UDS is an umbrella of service APIs provided by Unified CM. UDS provides a contact source API that can be used by Jabber over Cisco Expressway mobile and remote access for the contact source. The UDS contact source uses the Unified CM end user table information to provide a contact lookup service.

Beginning with Cisco Unified CM 11.5, the UDS-to-LDAP Proxy feature can also be used for contact searches. When enabled, contact searches are still handled by UDS but are proxied to the corporate LDAP directory, with UDS relaying results back to the Jabber client. This enables Jabber clients to search a corporate directory that exceeds the maximum number of users supported within Unified CM.

The UDS contact service is always used for remote Jabber clients connected over Expressway mobile and remote access, and it is an optional contact service for clients on the corporate network. You can populate the UDS contact source data by using the Unified CM web interface (by creating end users), by using the LDAP sync function to Active Directory or other supported LDAP source, or by using the UDS-to-LDAP Proxy function.

UDS does not support the same attribute list as LDAP contact sources. Rather, UDS supports the following attributes:

[username, firstname, lastname, middlename, nickname, phone number, homenumber, mobilenumber, email,directory URI, msURI, title,department, manager]

UDS does not provide the same level of predictive search or Ambiguous Name Resolution (ANR) provided by LDAP contact sources. UDS searches against firstname, lastname, and email address.

Note Ambiguous Name Resolution (ANR) is a search algorithm associated with Lightweight Directory Access Protocol (LDAP) clients, and it allows for objects to be bound without complex search filters. ANR is useful when you are locating objects and attributes that may or may not be known by the client.

UDS usage considerations for on-premise deployments:

- Jabber on-premises (on-net) can use LDAP or UDS as a contact source.
- UDS is a set of HTTP-based services provided by Unified CM. The UDS contact source is one UDS service, which provides contact and number resolution.
- When Jabber is connected remotely over Expressway mobile and remote access, it always uses UDS as a contact source. Again, your design must comply with cluster sizing recommendations.

UDS is an integrated network service that runs on all Unified CM servers in a cluster and is critical to server discovery. In the case of an IM-only deployment, although voice and video are not part of the deployment, Unified CM call processing subscriber pairs are still required to handle and distribute the UDS service load. The number of IM-only users will dictate the number of Unified CM subscriber pairs required. For example, a standard Unified CM cluster with 40,000 users and/or endpoints would require 4 subscriber pairs in order to support 40,000 IM-only users.

LDAP Directory

You can configure a corporate LDAP directory to satisfy a number of different requirements, including the following:

- User provisioning — You can provision users automatically from the LDAP directory into the Cisco Unified Communications Manager database using directory integration. Cisco Unified CM synchronizes with the LDAP directory content so that you avoid having to add, remove, or modify user information manually each time a change occurs in the LDAP directory.
- User authentication — You can authenticate users using the LDAP directory credentials. Cisco IM and Presence synchronizes all the user information from Cisco Unified Communications Manager to provide authentication for client users.
- User lookup — You can enable LDAP directory lookups to allow Cisco clients or third-party XMPP clients to search for contacts in the LDAP directory.

AD Groups and Enterprise Groups

Cisco Jabber users can search for groups in Microsoft Active Directory and add them to their contact lists.

Cisco Unified CM synchronizes its database with the Microsoft Active Directory groups at specified intervals (specified in LDAP Directory Synchronization Schedule parameters in the LDAP Directory configuration), and it also updates Jabber end-user contact lists upon synchronization.

If a Cisco Jabber user wants to add a group to the contact list while the AD Groups Sync feature is enabled, the Cisco Jabber client sends a group request to the IM and Presence Service node. The IM and Presence Service node provides the following information for each group member:

- Display Name
- User ID and Title
- Phone number
- Mail ID

AD Group Considerations for Groups and User Filters

- Group filters must be configured and validated by the administrator of the Cisco CallManager Application.
- The administrator should ensure that the User filters and Group filters are named appropriately and assigned to the right filter fields meant for User and Group filters.
- DirSync service does not validate the filter string format for syntax or logical accuracy; the administrator must verify the accuracy of the filters.
- For every filter string there should be one string added for ignoring security groups while performing synchronization.

WebEx Directory Integration

WebEx Directory Integration is achieved through the WebEx Administration Tool. WebEx imports a comma-separated value (CSV) file of your enterprise directory information into its WebEx Messenger service. For more information, refer to the documentation at

https://www.webex.com/webexconnect/orgadmin/help/index.htm?toc.htm?17444.htm

Directory Search

When a contact cannot be found in the local Jabber desktop client cache or contact list, a search for contacts can be made. The WebEx Messenger user can utilize a predictive search whereby the cache, contact list, and local Outlook contact list are queried as the contact name is being entered. If no matches are found, the search continues to query the corporate directory (WebEx Messenger database).

Common Deployment Models for Jabber Clients

Cisco Jabber desktop clients support the following deployment models:

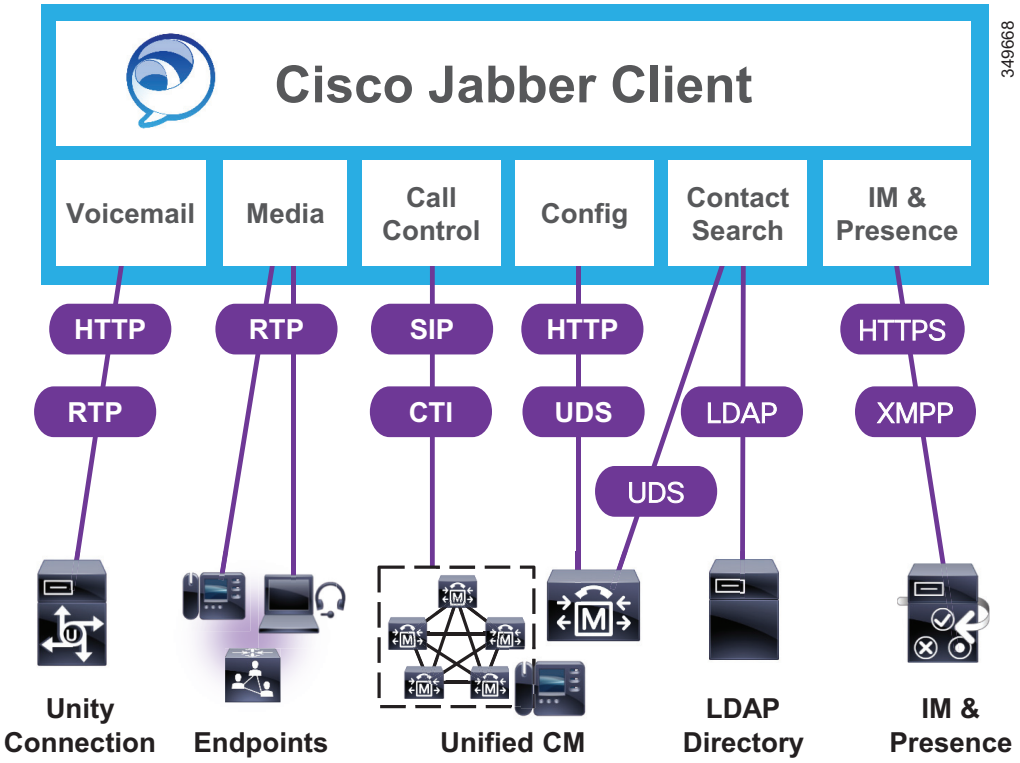
- On-Premises Deployment Model, page 20-11
- Cloud-Based Deployment Model, page 20-12
- Hybrid Cloud-Based and On-Premises Deployment Model, page 20-13
- Centralized IM and Presence Deployments, page 20-32

Your choice of deployment will depend primarily upon your product choice for IM and presence and the requirement for additional services such as voice and video, voicemail, and deskphone control.

On-Premises Deployment Model

In the on-premises deployment model, all services are set up and configured on an enterprise network that you manage and maintain. (See Figure 20-2.)

Figure 20-2 Jabber On-Premises Deployment Model



The on-premises deployment model for Cisco Jabber for Windows relies on the following components:

- Cisco Unified Communications Manager provides all user and device configuration capabilities.
- Cisco Unified Communications Manager and Cisco conferencing devices provide audio and video conferencing capabilities.
- Cisco Unity Connection provides voicemail capabilities.
- Cisco IM and Presence provides instant messaging and presence services.
- Microsoft Active Directory or another supported LDAP directory provides contact sources.

These components are the essential requirements to achieve a base deployment of Cisco Jabber clients. After you set up and configure a base deployment, you can set up and configure additional deployment options such as:

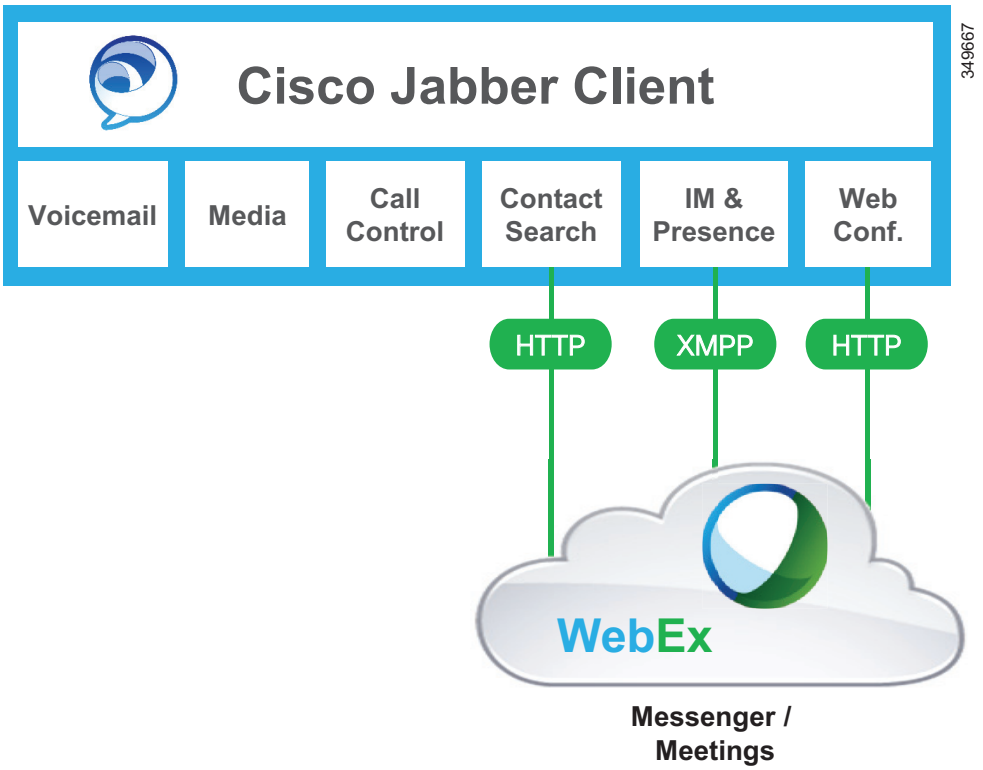
- Voice — Provides audio call capabilities.
- Video — Provides capabilities to enable users to transmit and receive video calls.
- Voicemail — Provides voicemail capabilities that users can retrieve directly in the Cisco Jabber client user interface or when users dial their voicemail number.

- Desktop sharing — Enables users to share their desktops via Binary Flow Control Protocol (BFCP).
- Microsoft Office integration — Provides user availability status and messaging capabilities directly through the user interface of Microsoft Office applications such as Microsoft Outlook.

Cloud-Based Deployment Model

In the cloud-based deployment model, all (or most) services are hosted in the cloud using Cisco WebEx. When implementing a cloud-based deployment model using Cisco WebEx, you manage and monitor your cloud-based deployment with the Cisco WebEx Administration Tool. (See Figure 20-3.)

Figure 20-3 Jabber Cloud-Based Deployment Model (WebEx)



The cloud-based deployment model for Cisco Jabber for Windows relies on Cisco WebEx Messenger service for the following services:

- Instant messaging and chat capabilities
- Presence capabilities for users
- Native desktop sharing
- User configuration and contact sources

These services are the essential components required to achieve a base deployment of Cisco Jabber for Windows. After you set up and configure a base deployment, you can set up and configure additional deployment options such as:

- Cisco WebEx Meeting Center — Offers hosted collaboration features such as online meetings and events.
- Microsoft Office integration — Provides user availability status and messaging capabilities directly through the user interface of Microsoft Office applications such as Microsoft Outlook. This integration is set up by default.
- Calendar integration — Calendar integration with WebEx Meeting Center, Outlook, and IBM Lotus Notes is also supported.

For information on WebEx Messenger service configuration for Jabber Clients, refer to the *Cisco WebEx Messenger Administrator's Guide*, available at

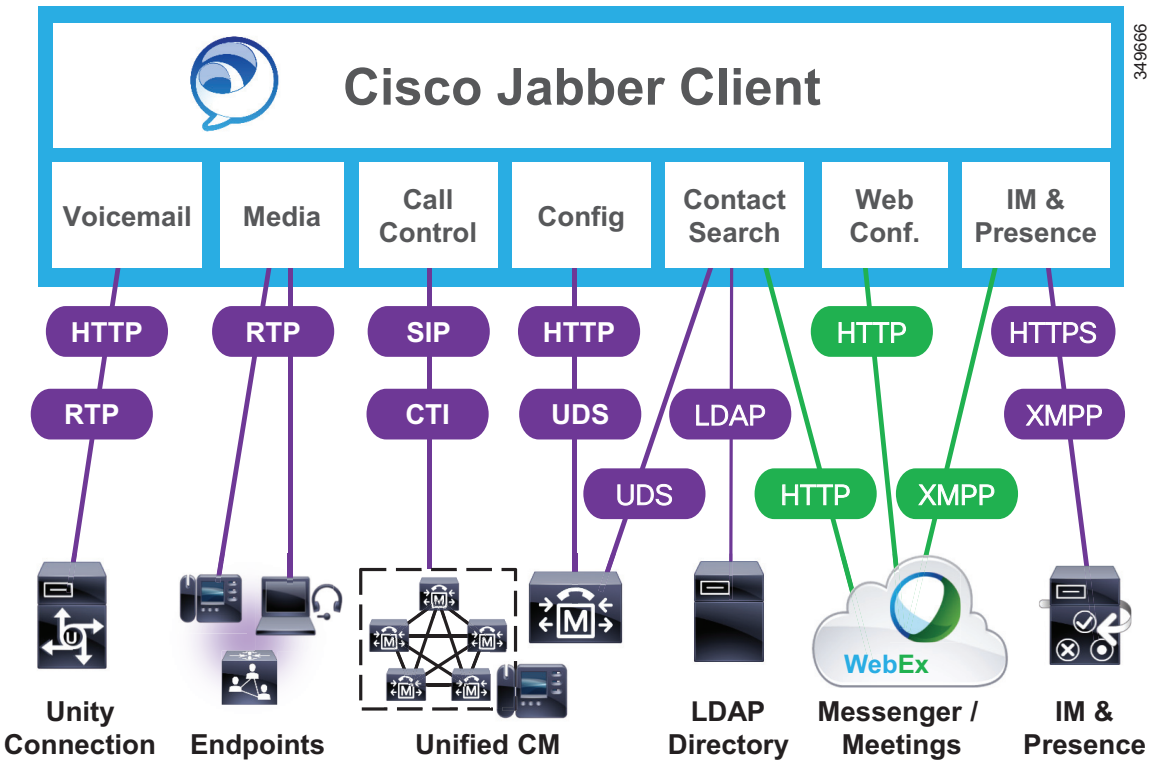
https://www.webex.com/webexconnect/orgadmin/help/index.htm

Hybrid Cloud-Based and On-Premises Deployment Model

In a hybrid deployment, the cloud-base services hosted on Cisco WebEx Messenger service are combined with the following components of an on-premises deployment (see Figure 20-4):

- Cisco Unified Communications Manager provides user and device services.
- Cisco Unity Connection provides voicemail services.

Figure 20-4 Jabber Hybrid Cloud-Based and On-Premises Deployment Model



Client-Specific Design Considerations

The following sections discuss design considerations that are specific to Cisco Collaboration desktop clients for Mac and Windows. For common design considerations for these client types, refer to the design guidance provided in the section on Cisco Jabber Desktop Client Architecture, page 8-23.

Phone-Specific Presence and Busy Lamp Field

Endpoints connected to Unified CM can receive the line status of one or more other endpoints as idle, busy, unknown. The status is shown in the call history, in the directory, and by the use of the busy lamp field (BLF) feature. While presence on the call history and directory is received only after a lookup performed by the user, BLF constantly monitors the line status of a telephone or a video phone and represents it on a specific presence-enabled speed-dial configured in the monitoring device.

All telephony presence requests for users, whether inside or outside the cluster, are processed and handled by Cisco Unified CM.

A Unified CM watcher that sends a presence request will receive a direct response, including the presence status, if the watcher and presence entity are within the same Unified CM cluster.

If the presence entity exists outside the cluster, Unified CM will query the external presence entity through the SIP trunk. If the watcher has permission to monitor the external presence entity based on the SUBSCRIBE calling search space and presence group (both described in the section on Unified CM Presence Policy, page 20-17), the SIP trunk will forward the presence request to the external presence entity, await the presence response from the external presence entity, and return the current presence status to the watcher.

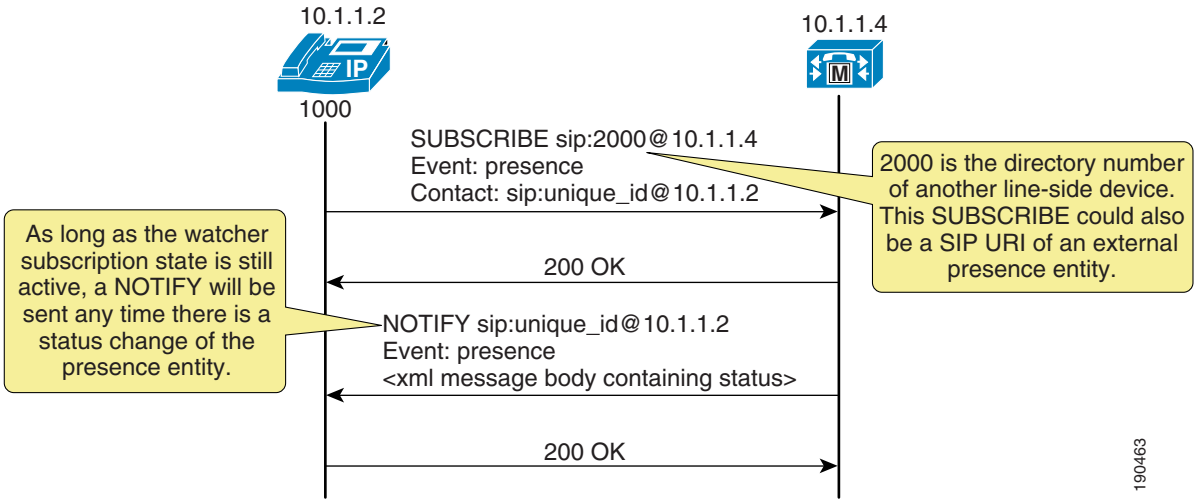
A watcher that is not in a Unified CM cluster can send a presence request to a SIP trunk. If Unified CM supports the presence entity, it will respond with the current presence status. If Unified CM does not support the presence entity, it will reject the presence request with a SIP error response.

Unified CM Presence with SIP

Unified CM uses the term *SIP line* to represent endpoints supporting SIP that are directly connected and registered to Unified CM and the term *SIP trunk* to represent trunks supporting SIP. SIP line-side endpoints acting as presence watchers can send a SIP SUBSCRIBE message to Unified CM requesting the presence status of the indicated presence entity.

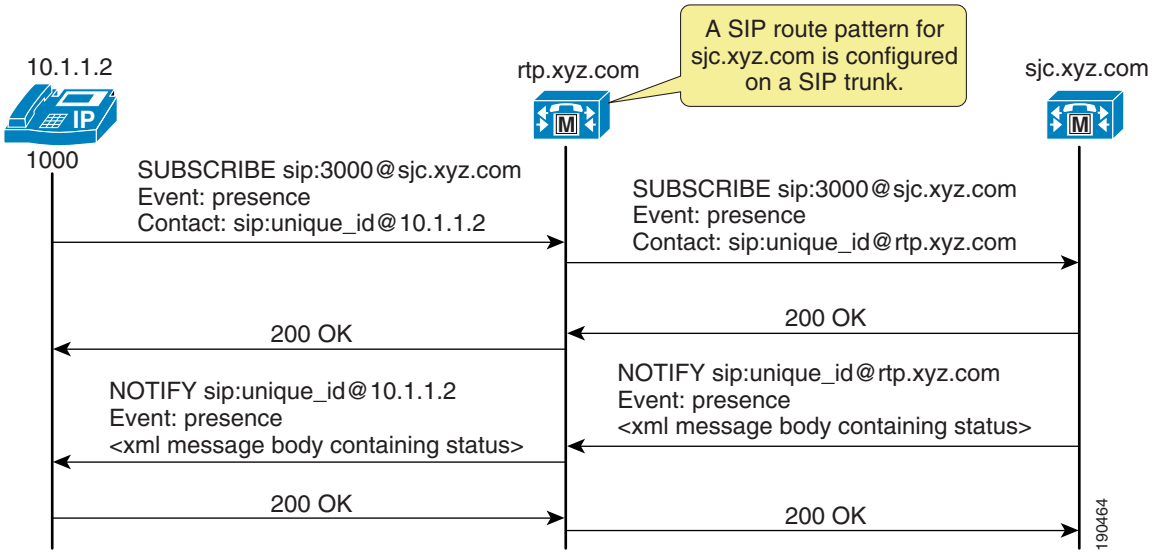
If the presence entity resides within the Unified CM cluster, Unified CM responds to a SIP line-side presence request by sending a SIP NOTIFY message to the presence watcher, indicating the current status of the presence entity. (See Figure 20-5.)

Figure 20-5 SIP Line SUBSCRIBE/NOTIFY Exchange



If the presence entity resides outside the Unified CM cluster, Unified CM routes a SUBSCRIBE request out the appropriate SIP trunk, based on the SUBSCRIBE calling search space, presence group, and SIP route pattern. When Unified CM receives a SIP NOTIFY response on the trunk, indicating the presence entity status, it responds to the SIP line-side presence request by sending a SIP NOTIFY message to the presence watcher, indicating the current status of the presence entity. (See Figure 20-6.)

Figure 20-6 SIP Trunk SUBSCRIBE/NOTIFY Exchange



SUBSCRIBE messages for any directory number or SIP URI residing outside the Unified CM cluster are sent or received on a SIP trunk in Unified CM. The SIP trunk could be an interface to another Unified CM or it could be an interface to the Cisco IM and Presence Service.

Unified CM Speed Dial Presence

Unified CM supports the ability for a speed dial to have presence capabilities by means of a busy lamp field (BLF) speed dial. BLF speed dials work as both a speed dial and a presence indicator. However, only the system administrator can configure a BLF speed dial; a system user is not allowed to configure a BLF speed dial.

The administrator must configure the BLF speed dial with a target directory number or URI that is resolvable to a directory number or URI within the Unified CM cluster or a SIP trunk destination. BLF SIP line-side endpoints can also be configured with a SIP URI for the BLF speed dial, but SCCP line-side endpoints cannot be configured with a SIP URI for BLF speed dial. The BLF speed dial indication is a line-level indication and not a device-level indication.

Note A maximum of 30,000 BLF speed dials can be configured per cluster.

For a listing of the phone models that support BLF speed dials, consult the Cisco Unified IP Phone administration guides available on https://www.cisco.com/.

Figure 20-7 lists the various types of BLF speed dial indications from the phones.

Figure 20-7 Indicators for Speed Dial Presence on Cisco Unified IP Phones 7900 Series

State	Icon	LED
Idle		
Busy		
Unknown		

190465

Unified CM Call History Presence

Unified CM supports presence capabilities for call history lists (the Directories button on the phone). Call history list presence capabilities are controlled via the **BLF for Call Lists** Enterprise Parameter within Unified CM Administration. The **BLF for Call Lists** Enterprise Parameter impacts all pages using the phone Directories button (Missed, Received, and Placed Calls, Personal Directory, or Corporate Directory), and it is set on a global basis.

For a listing of the phone models that support presence capabilities for call history lists, consult the Cisco Unified IP Phone administration guides available on https://www.cisco.com/.

The presence indicators for call history lists are the same as those listed in the Icon column in Figure 20-7; however, no LED indications are available.

Unified CM Presence Policy

Unified CM provides the capability to set policy for users who request presence status. You can set this policy by configuring a calling search space specifically to route SIP SUBSCRIBE messages for presence status and by configuring presence groups with which users can be associated to specify rules for viewing the presence status of users associated with another group.

Unified CM Subscribe Calling Search Space

The first aspect of presence policy for Unified CM is the SUBSCRIBE calling search space. Unified CM uses the SUBSCRIBE calling search space to determine how to route presence requests (SUBSCRIBE messages with the Event field set to Presence) that come from the watcher, which could be a phone or a trunk. The SUBSCRIBE calling search space is associated with the watcher and lists the partitions that the watcher is allowed to "see." This mechanism provides an additional level of granularity for the presence SUBSCRIBE requests to be routed independently from the normal call-processing calling search space.

The SUBSCRIBE calling search space can be assigned on a device basis or on a user basis. The user setting applies for originating subscriptions when the user is logged in to the device through Extension Mobility or when the user is administratively assigned to the device.

With the SUBSCRIBE calling search space set to <None>, BLF speed dial and call history list presence status does not work and the subscription messages is rejected as “user unknown.” When a valid SUBSCRIBE calling search space is specified, the indicators work and the SUBSCRIBE messages are accepted and routed properly.

Note Cisco strongly recommends that you do not leave any calling search space defined as <None>. Leaving a calling search space set to <None> can introduce presence status or dialing plan behavior that is difficult to predict.

Unified CM Presence Groups

The second aspect of the presence policy for Unified CM is presence groups. Devices, directory numbers, and users can be assigned to a presence group, and by default all users are assigned to the Standard Presence Group. A presence group controls the destinations that a watcher can monitor, based on the user’s association with their defined presence group (for example, Contractors watching Executives is disallowed, but Executives watching Contractors is allowed). The presence group user setting applies for originating subscriptions when the user is logged in to the device via Extension Mobility or when the user is administratively assigned to the device.

When multiple presence groups are defined, the Inter-Presence Group Subscribe Policy service parameter is used. If one group has a relationship to another group via the Use System Default setting rather than being allowed or disallowed, this service parameter's value will take effect. If the Inter-Presence Group Subscribe Policy service parameter is set to **Disallowed**, Unified CM will block the request even if the SUBSCRIBE calling search space allows it. The Inter-Presence Group Subscribe Policy service parameter applies only for presence status with call history lists and is not used for BLF speed dials.

Presence groups can list all associated directory numbers, users, and devices if you enable dependency records. Dependency records allow the administrator to find specific information about group-level settings. However, use caution when enabling the Dependency Record Enterprise parameter because it could lead to high CPU usage.

Unified CM Presence Guidelines

Unified CM enables the system administrator to configure and control user phone state presence capabilities from within Unified CM Administration. Observe the following guidelines when configuring presence within Unified CM:

- Select the appropriate model of Cisco Unified IP Phones that have the ability to display user phone state presence status.
- Define a presence policy for presence users.
  - Use SUBSCRIBE calling search spaces to control the routing of a watcher presence-based SIP SUBSCRIBE message to the correct destinations.
  - Use presence groups to define sets of similar users and to define whether presence status updates of other user groups are allowed or disallowed.
- Call history list presence capabilities are enabled on a global basis; however, user status can be secured by using a presence policy.
- BLF speed dials are administratively controlled and are not impacted by the presence policy configuration.

Note Cisco Business Edition can be used in ways similar to Unified CM to configure and control user presence capabilities. For more information, refer to the chapter on Call Processing, page 9-1.

User Presence: Cisco IM and Presence Architecture

The Cisco IM and Presence Service uses standards-based XMPP for instant messaging and presence. The Cisco IM and Presence Service also supports SIP for interoperability with SIP IM providers. Cisco IM and Presence also provides an HTTP interface that has a configuration interface through Simple Object Access Protocol (SOAP); a presence interface through Representational State Transfer (REST); and a presence, instant messaging, and Bidirectional-streams Over Synchronous HTTP (BOSH) interface through the Cisco AJAX XMPP Library (CAXL). The Cisco AJAX XMPP Library web tool kit communicates to the BOSH interface on the Extensible Communications Platform within Cisco IM and Presence. The Cisco IM and Presence Service collects, aggregates, and distributes user capabilities and attributes using these standards-based SIP, SIMPLE, XMPP, and HTTP interfaces.

Cisco or third-party applications can integrate with presence and provide services that improve the end-user experience and efficiency. The core components of the Cisco IM and Presence Service consist of: the Extensible Communications Platform (XCP), which handles presence, instant messaging, roster, routing, policy, and federation management; the Rich Presence Service, which handles presence state gathering, network-based rich presence composition, and presence-enabled routing functionality; and support for ad-hoc group chat storage with persistent chat and message archiving handled to an external database. If persistent chat is enabled, ad-hoc rooms are stored to the external PostgreSQL, Microsoft SQL, or Oracle database for the duration of the ad-hoc chat. If persistent chat is disabled, ad-hoc chats are stored in volatile memory for the duration of the chat.

Applications (either Cisco or third-party) can integrate presence and provide services that improve the end user experience and efficiency. In addition, Cisco Jabber is a supported client of the Cisco IM and Presence Service that also integrates instant messaging and presence status.

The Cisco IM and Presence Service also contains support for interoperability with Microsoft Lync Server 2010 and 2013 and the Microsoft Lync client for any Cisco Unified IP Phone connected to a Unified CM. The Microsoft Lync client interoperability includes click-to-dial functionality, phone control capability through Remote Call Control (RCC), and presence status of Cisco Unified IP Phones.

On-Premises Cisco IM and Presence Service Cluster

The Cisco IM and Presence Service uses the same underlying appliance model and hardware used by Unified CM as well as Unified CM on the Cisco Unified Computing System (UCS) platform, including a similar administration interface. For details on the supported platforms, refer to the latest version of the *Cisco Unified Communications Manager Compatibility Matrix*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps556/products_device_support_tables_list.html

A Cisco IM and Presence Service cluster consists of up to six servers, including one designated as a publisher, which utilize the same architectural concepts as the Unified CM publisher and subscriber. Within a Cisco IM and Presence Service cluster, individual servers can be grouped to form a subcluster, and the subcluster can have at most two servers associated with it. Figure 20-8 shows the basic topology for a Cisco IM and Presence Service cluster, while Figure 20-9 shows a highly available topology. The Cisco IM and Presence Service cluster can also have mixed subclusters, where one subcluster is configured with two servers while other subclusters contain a single server, as shown in Figure 20-10.

Figure 20-8 Basic Deployment of On-Premises Cisco IM and Presence Service

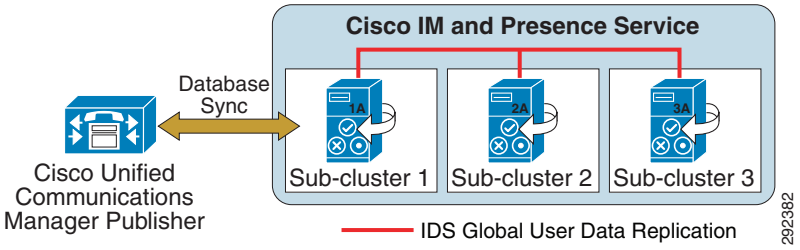


Figure 20-9 High Availability Deployment of On-Premises Cisco IM and Presence Service

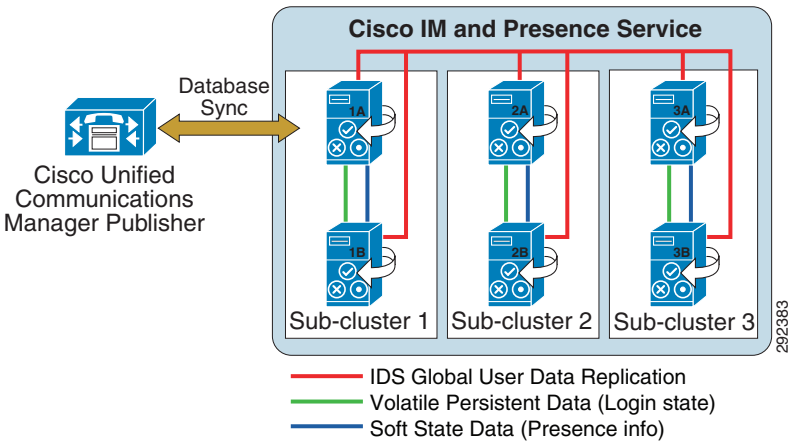
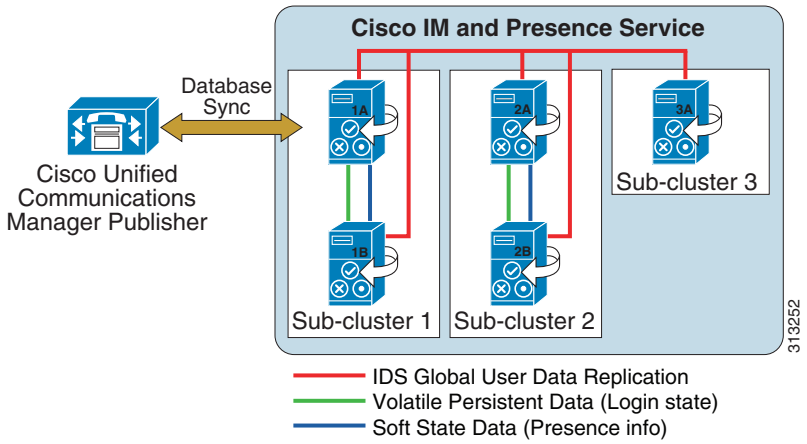


Figure 20-10 Mixed Deployment of On-Premises Cisco IM and Presence Service



The on-premises Cisco IM and Presence Service utilizes and builds upon the database used by the Unified CM publisher by sharing the user and device information.

To support Availability Integration with Cisco Unified CM, the CUCM Domain SIP Proxy service parameter must match the DNS domain of the Unified CM cluster. By default, the CUCM Domain SIP Proxy service parameter is set to the DNS domain of the IM and Presence database publisher node. Therefore, if the DNS domain of the IM and Presence database publisher node differs from the DNS domain of the Unified CM cluster, you must update this service parameter using the Cisco Unified CM IM and Presence Administration user interface on the IM and Presence database publisher node. For more information on specifying the DNS domain associated with the Cisco Unified Communications Manager cluster, refer to the latest version of *Configuration and Administration of IM and Presence Service on Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Keep in mind that IM and Presence maintains access control list (ACL) entries for all Unified CM nodes. These entries are FQDN-based and are generated by appending the Unified CM hostname to the DNS domain of the IM and Presence publisher. So if the DNS domain of IM and Presence publisher (database) is different than that of the Unified CM cluster, invalid ACL entries will be generated, which would cause issues.

Note A single Unified CM cluster supports only a single IM and Presence Service cluster; therefore, a separate IM and Presence Service cluster is required for each Unified CM cluster.

Intracluster traffic participates at a very low level between the Cisco IM and Presence Service and Unified CM, and between the Cisco IM and Presence Service publisher and subscriber servers. Both clusters share a common hosts file and have a strong trust relationship using IPTables. At the level of the database and services, the clusters are separate and distinct; however, the configuration and administration is primarily done on the Unified CM cluster, with limited configuration and administration done on the IM and Presence Service cluster. There is currently no Transport Layer Security (TLS) or IPSec utilization for intracluster traffic.

The Cisco IM and Presence Service publisher communicates directly with the Unified CM publisher via the AVVID XML Layer Application Program Interface (AXL API) using the Simple Object Access Protocol (SOAP) interface. When first configured, the Cisco IM and Presence Service publisher performs an initial synchronization of the entire Unified CM user and device database. All Cisco IM and Presence Service users are configured in the Unified CM End User configuration. During the synchronization, Cisco IM and Presence Service populates these users in its database from the Unified CM database and does not provide end-user configuration from its administration interface. After synchronization, users must be enabled for IM and Presence Service through the Cisco Unified Communications Manager administrator interface before the Cisco IM and Presence Service can mange them.

Note Cisco IM and Presence Service supports synchronization of up to 160,000 users, equivalent to Unified CM. However, the maximum number of licensed presence users for a Cisco IM and Presence Service cluster is 75,000 using the 25k-user IM and Presence VM template across 3 sub-cluster pairs, assuming a megacluster deployment.

The initial Cisco IM and Presence Service database synchronization from Unified CM might take a while, depending on the amount of information in the database as well as the load that is currently on the system. Subsequent database synchronizations from Unified CM to Cisco IM and Presence Service are performed in real time when any new user or device information is added to Unified CM.

Note When Cisco IM and Presence Service is performing the initial database synchronization from Unified CM, do not perform any administrative activities on Unified CM while the synchronization agent is active.

On-Premises Cisco IM and Presence Service High Availability

The Cisco IM and Presence Service cluster consists of up to a maximum of six servers, which can be configured into multiple subclusters with the maximum of three subclusters for high availability. A subcluster contains a maximum of two servers and allows for users associated with one server of the subcluster to use the other server in the subcluster automatically if a failover event occurs. Cisco IM and Presence Service does not provide failover between subclusters.

When deploying a Cisco IM and Presence Service cluster for high availability, you must take into consideration the maximum number of users per server to avoid oversubscribing any one server within the subcluster in the event of a failover.

You can achieve high availability using two different setups: active/active or active/standby. You can set up the nodes in a presence redundancy group to work together in Balanced Mode, which provides redundant high availability with automatic user load balancing and user failover in case one of the nodes fails because of component failure or power outage. In an active/standby setup, the standby node automatically takes over for the active node if the active node fails.

Deployment Consideration

The IM and Presence Service cluster supports a maximum of 75,000 users in Full UC mode, across 3 single IM and Presence nodes deployed with the 25,000-user VM configuration template but with no high availability.

A high availability deployment for 75,000 users would require 3 IM and Presence subcluster pairs deployed with the 25,000-user VM configuration template option.

Note The 25,000-user VM configuration should be used for megacluster deployments, which must be reviewed and approved by Cisco prior to deployment. If your deployment is not a megacluster, then use a lower capacity VM configuration template, such as the 15,000-user or 5,000-user VM configuration template. However, if it is more desirable to use the 25k-user IM and Presence VM configuration in your non-megacluster deployment, contact Cisco with your design topology and request approval prior to deploying the 25k-user VM configuration. Exceptions may be provided based on a design review of the IM and Presence configuration using the 25k-user VM template.

On-Premises Cisco IM and Presence Service Deployment Models

Unified CM provides a choice of the following deployment models:

- Single site
- Multisite WAN with centralized call processing
- Multisite WAN with distributed call processing
- Clustering over the WAN
- Centralized IM and Presence

Cisco IM and Presence Service is supported with all the Unified CM deployment models. However, Cisco recommends locating the Cisco IM and Presence Service publisher in the same physical datacenter as the Unified CM publisher due to the initial user database synchronization. All on-premises Cisco IM and Presence servers should be physically located in the same datacenter within the Cisco IM and Presence Service cluster, with the exception of geographic datacenter redundancy and clustering over the WAN (for details, see Clustering Over the WAN, page 20-29).

For more information on Unified CM deployment models, see the chapter on Collaboration Deployment Models, page 10-1.

Cisco IM and Presence Service deployment depends on high-availability requirements, the total number of users, and the server being used. Detailed configuration and deployment steps can be found in the *Deployment Guide for Cisco IM and Presence*, available at

https://www.cisco.com/en/US/products/ps6837/products_installation_and_configuration_guides_list.html

A highly available Cisco IM and Presence Service cluster requires two servers per subcluster. This allows for users to fail-over between the servers within the subcluster; however, the total number of users supported and the time to failover vary based on which features are enabled, the average size of contact lists, the rate of traffic placed on the servers, and the placement of the servers if deployed across a WAN. Once a Cisco IM and Presence Service subcluster is configured for two servers, it always operates as highly available if **High Availability** is configured in the Unified CM administration **System > Presence Redundancy Group**. High availability can be deployed using an Active/Standby model or an Active/Active model, and these modes are controlled by the Enterprise Parameter **User Assignment Mode for Presence Server**. By default all users are balanced across all servers in the cluster, and Cisco recommends leaving this parameter set to its default value.

Note Each subcluster is a Presence Redundancy Group.

Cisco IM and Presence Active/Standby mode — (Setting **User Assignment Mode for Presence Server** to **None**) is attained by manually assigning users to the first server in the subcluster, leaving the second server with no users assigned but all processes synchronized and ready for a failover if the first server in

the subcluster fails. For example, in Figure 20-9 the first user would be assigned to server 1A, the second user to server 2A, the third user to server 3A, the fourth user to server 1A, the fifth user to server 2A, the sixth user to server 3A, and so forth. The users should be assigned equally across all the 'A' servers in the cluster.

Cisco IM and Presence Active/Active mode — (Setting **User Assignment Mode for Presence Server** to **balanced**), which is the default and recommended setting for load distribution, will automatically assign users equally across all servers in the subclusters. Each server is synchronized and ready for a failover if the other server in the subcluster fails. For example, in Figure 20-9 the first user would be assigned to server 1A, the second user to server 2A, the third user to server 3A, the fourth user to server 1B, the fifth user to server 2B, the sixth user to server 3B, and so forth. The users are assigned equally across all the servers in the cluster.

Cisco IM and Presence Active/Active deployments with a balanced **User Assignment Mode for Presence Server** allows for redundancy flexibility based on the features being used, the size of user contact lists, and the traffic (user data profiles) being generated. A Cisco IM and Presence Active/Active deployment with a fully redundant mode, regardless of features, requires the total number of supported users to be reduced in half (for example, a deployment of 15,000 Users OVAs in a balanced high-availability redundant configuration supports up to 15,000 users per subcluster). A Cisco IM and Presence Active/Active deployment with a non-redundant mode requires a more detailed look at the Cisco IM and Presence Service features being utilized, the average size of the users contact lists, as well as the traffic being generated. For example, for a deployment with presence and instant messaging enabled and calendaring and mobility integration disabled, with an average contact list of 30 users and a user data profile of a few presence and instant message updates, it is possible to support more than 15,000 users per subcluster.

A Cisco IM and Presence Service cluster deployment that is not highly available allows each server in the subcluster to support up to the maximum number of users for the server, and the total number of supported users for all servers in the cluster can be up to the maximum number of users for the IM and Presence Service cluster. Once a second server is added in a subcluster, the subcluster will still act as if in a high-available deployment; however, if a server failure occurs, an attempt to fail-over might not result in success if the online server reaches its capacity limit based on the Cisco IM and Presence Service features enabled, the average user contact list size, and the traffic being generated by the users.

On-Premises Cisco IM and Presence Service Deployment Examples

Example 20-1 Single Unified CM Cluster with Cisco IM and Presence Service

Deployment requirements:

- 4,000 users that could scale up to 13,000 users
- Single Cisco Unified Communications Manager cluster
- Instant message logging and compliance are not needed
- High availability is not needed

Hardware and software platform:

- Cisco UCS virtual machine for one 15,000-user VM configuration Full UC template

Deployment:

- One single-server subcluster using User Assignment Mode for Presence Server = balanced

Example 20-2 Two Unified CM Clusters with Cisco IM and Presence Service

Deployment requirements:

- 11,000 users that could scale up to 24,000 users
- Two Cisco Unified Communications Manager clusters
- Instant message logging and compliance are not needed
- High availability is not needed

Hardware and software platform:

- Cisco UCS virtual machines for two 15,000-user VM configuration Full UC templates

Deployment:

- Two Cisco IM and Presence Service clusters (one per Cisco Unified Communications Manager cluster), each with one server using User Assignment Mode for Presence Server = balanced

Example 20-3 Single Unified CM Cluster with Cisco IM and Presence Service

Deployment requirements:

- 500 users that could scale up to 2500 users
- Single Cisco Unified Communications Manager cluster
- Instant message archiving is required
- High availability is required

Hardware:

- Cisco UCS virtual machines for two 5,000-user VM configuration Full UC templates

Deployment:

- One two-server subcluster using User Assignment Mode for Presence Server set to **balanced**, with a PostgreSQL, Microsoft SQL, or Oracle database instance for the cluster

Example 20-4 Single Cisco Business Edition Cluster with Cisco IM and Presence Service

Deployment requirements:

- 150 users that could scale up to 1,000 users
- Single Cisco Business Edition
- Instant message archiving and persistent chat are required
- High availability is required

Hardware:

- Cisco Business Edition 6000H using the 1,000-user VM configuration Full UC template

Deployment:

- One two-server subcluster using User Assignment Mode for Presence Server set to **balanced**, with a unique PostgreSQL, Microsoft SQL, or Oracle database instance per server in the cluster for persistent chat functionality

Example 20-5 Megacluster with Cisco IM and Presence Service

This is just an example of what megacluster deployment might consist of, if needed to deploy more than 40,000 users and/or devices.

Note All deployments requiring more than 4 Unified CM subscriber pairs must be reviewed and approved by the Cisco Megacluster team prior to deployment. For more details on the megacluster approval and review process, refer to the information at https://wiki.cisco.com/display/CCM/Megacluster.

IM and Presence deployment requirements:

- Six nodes using the 25,000-user IM and Presence VM configuration template
- Node distributed as 3 subcluster subscriber pairs in high availability mode

Unified CM deployment requirements:

- Nineteen nodes using 10,000-user Unified CM VM configuration template
- Unified CM deployed with dedicated publisher
- Dedicated TFTP1 server and backup TFTP2 server
- Eight subscriber pair virtual nodes
- Instant message compliance is required
- High availability is required

Hardware platform:

- Cisco UCS B-Series with Unified CM 10,000-user VM configuration template and 25,000-user IM and Presence VM configuration template

Example 20-6 Multiple Unified CM Clusters with Cisco IM and Presence Service on a Single UCS B-Series Server

Deployment requirements:

- 75,000 users belonging to five different Unified CM clusters
- Instant message compliance is required
- High availability is required

Hardware and software platform:

- Cisco UCS B-Series with ten 15,000-user VM configuration Full UC templates

Deployment

- Five Cisco IM and Presence Service clusters in a single platform UCS B-Series, each one serving one of the five Unified CM clusters with 15,000 users each.

On-Premises Cisco IM and Presence Service Performance

Cisco IM and Presence Service clusters support single-server as well as multi-server configurations. The maximum number of users supported for a Cisco IM and Presence Service cluster is based on the platform being used in the deployment. For example, if a Cisco IM and Presence Service cluster is deployed with three 2,000-user VM configuration templates, each forming their own subcluster, then a total of 6,000 users would be supported. For a Cisco IM and Presence Service cluster, the maximum number of full UC users supported is 75,000 and cannot exceed the maximum number of supported devices for cluster deployments. The maximum number of IM-only users supported is 75,000. For a complete list of platform requirements for the Cisco IM and Presence Service, as well as the maximum number of users supported per platform, refer to the documentation available at

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-cisco-ucm-im-presence.html

The Cisco IM and Presence Service supports deployments using only VM configuration templates on virtualized servers; physical server support is not available for the Cisco IM and Presence Service.

Cisco recommends using identical VM configurations for all IM and Presence nodes in a cluster. However, mixing VM configurations of different capacities within a cluster is allowed as long as the VM configuration used for the IM and Presence publisher node is of equal capacity or larger than the VM configuration used on any of the subscriber nodes in the same cluster.

Similar guidelines apply to redundancy and high availability models. Within a cluster, the VM configuration used on the IM and Presence standby and/or backup virtual subscriber node must not be larger than the VM configuration of the IM and Presence publisher or its own respective active and/or primary subscriber.

If you are mixing VM configurations within the same IM and Presence cluster, consider the possible impact to performance and capacity due to the various VM configurations being used. The overall cluster capacity might ultimately be dictated by the capacity of the smallest VM configuration within the cluster.

Note IM and Presence integration with Unified CM does not imply that they are part of the same cluster but rather two separate clusters.

On-Premises Cisco IM and Presence Service Deployment

Cisco IM and Presence Service can be deployed in any of the following configurations:

- Single-Cluster Deployment, page 20-26
- Intercluster Deployment, page 20-29
- Clustering Over the WAN, page 20-29
- Federated Deployment, page 20-36

Single-Cluster Deployment

Figure 20-11 represents the communication protocols between the Cisco IM and Presence Service, the LDAP server, and Cisco Unified Communications Manager for basic functionality. For complete information on Cisco IM and Presence Service administration and configuration, refer to the Cisco IM and Presence installation, administration, and configuration guides, available at

https://www.cisco.com/en/US/products/ps6837/tsd_products_support_series_home.html

Figure 20-11 Interactions Between Cisco IM and Presence Service Components

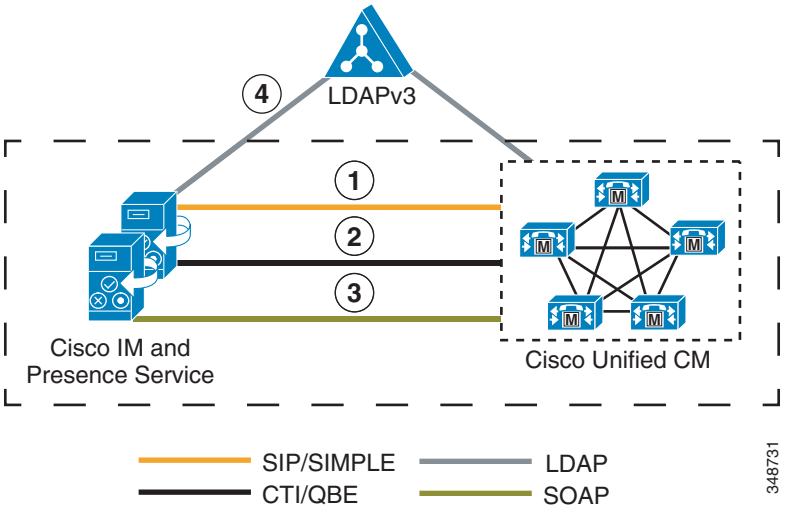


Figure 20-11 depicts the following interactions between Cisco IM and Presence Service components:

1. The SIP connection between the Cisco IM and Presence Service and Unified CM handles all the phone state presence information exchange.

   Unified CM configuration requires the Cisco IM and Presence Service to be added as application servers on Unified CM and also requires a SIP trunk pointing to the Cisco IM and Presence Service. The address configured on the SIP trunk could be a Domain Name System (DNS) server (SRV) fully qualified domain name (FQDN) that resolves to the Cisco IM and Presence Services, or it could simply be an IP address of an individual Cisco IM and Presence Service. The Cisco IM and Presence Service handles the configuration of the Cisco Unified Communications Manager application server entry automatically through AXL/SOAP once the administrator adds a node in the system topology page through Cisco IM and Presence Service administration.

   If DNS is highly available within your network and DNS SRV is an option, configure the SIP trunk on Unified CM with a DNS SRV FQDN of the Cisco IM and Presence Service publisher and subscriber. Also configure the Presence Gateway on the Cisco IM and Presence Service with a DNS SRV FQDN of the Unified CM subscribers, equally weighted. This configuration will allow for presence messaging to be shared equally among all the servers used for presence information exchange.

   If DNS is not highly available or not a viable option within your network, use IP addressing. When using an IP address, presence messaging traffic cannot be equally shared across multiple Unified CM subscribers because it points to a single subscriber.

   Unified CM provides the ability to further streamline communications and reduce bandwidth utilization by means of the service parameter IMP PUBLISH Trunk, which allows for the PUBLISH method (rather than SUBSCRIBE/NOTIFY) to be configured and used on the SIP trunk interface to Cisco IM and Presence Service. Once the IMP PUBLISH Trunk service parameter has been enabled, the users must be associated with a line appearance and not just a primary extension.

2. The Computer Telephony Integration Quick Buffer Encoding (CTI-QBE) connection between Cisco IM and Presence Service and Unified CM is the protocol used by presence-enabled users in Cisco IM and Presence Service to control their associated phones registered to Unified CM. This CTI communication occurs when Cisco Jabber is using Desk Phone mode to do Click to Call or when Microsoft Office Communicator is doing Click to Call through Microsoft Office Communications Server 2007 or Microsoft Lync.

   a. Unified CM configuration requires the user to be associated with a CTI Enabled Group, and the primary extension assigned to that user must be enabled for CTI control (checkbox on the Directory Number page). The CTI Manager Service must also be activated on each of the Unified CM subscribers used for communication with the Cisco IM and Presence Service publisher and subscriber. Integration with Microsoft Office Communications Server 2007 or Microsoft Lync requires that you configure an Application User, with CTI Enabled Group and Role, on Unified CM.

   b. Cisco IM and Presence Service CTI configuration (CTI Server and Profile) for use with Cisco Jabber is automatically created during the database synchronization with Unified CM. All Cisco Jabber CTI communication occurs directly with Unified CM and not through the Cisco IM and Presence Service.

      Cisco IM and Presence Service CTI configuration (Desktop Control Gateway) for use with Microsoft Office Communications Server 2007 or Microsoft Lync requires you to set the Desktop Control Gateway address (Cisco Unified Communications Manager Address) and a provider, which is the application user configured previously in Unified CM. Up to eight Cisco Unified Communications Manager Addresses can be provisioned for increased scalability. Only IP addresses can be used for Desktop Control Gateway configuration in the Cisco IM and Presence Service. Administrators should ensure that any configuration and assignment of Cisco Unified Communications Manager addresses is evenly distributed for the purpose of load balancing.

3. The AXL/SOAP interface handles the database synchronization from Unified CM to populate the Cisco IM and Presence Service database.

   a. No additional configuration is required on Unified CM.

   b. Cisco IM and Presence Service security configuration requires you to set a user and password for the Unified CM AXL account in the AXL configuration.

      The Sync Agent Service Parameter, User Assignment, set to **balanced** by default, will load-balance all users equally across all servers within the Cisco IM and Presence Service cluster. The administrator can also manually assign users to a particular server in the Cisco IM and Presence Service cluster by changing the User Assignment service parameter to **None**.

4. The LDAP interface is used for LDAP authentication of users. For more information regarding LDAP synchronization and authentication, see the chapter on Directory Integration and Identity Management, page 16-1.

   Unified CM is responsible for all user entries via manual configuration or synchronization directly from LDAP, and Cisco IM and Presence Service then synchronizes all the user information from Unified CM. If a user logs into the Cisco IM and Presence Service and LDAP authentication is enabled on Unified CM, Cisco IM and Presence Service will go directly to LDAP for the user authentication using the Bind operation.

   When using Microsoft Active Directory, consider the choice of parameters carefully. Performance of Cisco IM and Presence Service might be unacceptable when a large Active Directory implementation exists and the configuration uses a Domain Controller. To improve the response time of Active Directory, it might be necessary to promote the Domain Controller to a Global Catalog and configure the LDAP port as 3268.

Intercluster Deployment

The deployment topology in previous sections is for a single Cisco IM and Presence Service cluster communicating with a single Unified CM cluster. Presence and instant messaging functionality is limited by having communications within a single cluster only. Therefore, to extend presence and instant messaging capability and functionality, these standalone clusters can be configured for peer relationships for communication between clusters within the same domain. This functionality provides the ability for users in one cluster to communicate and subscribe to the presence of users in a different cluster within the same domain.

To create a fully meshed presence topology, each Cisco IM and Presence Service cluster requires a separate peer relationship for each of the other Cisco IM and Presence Service clusters within the same domain. The address configured in this intercluster peer could be a DNS FQDN that resolves to the remote Cisco IM and Presence Service cluster servers, or it could also simply be the IP address of the Cisco IM and Presence Service cluster servers.

The interface between each Cisco IM and Presence Service cluster is two-fold, an AXL/SOAP interface and a signaling protocol interface (SIP or XMPP). The AXL/SOAP interface, between publisher-only servers of an IM and Presence Service cluster, handles the synchronization of user information for home cluster association, but it is not a full user synchronization. The signaling protocol interface (SIP or XMPP) is a full mesh between all servers within the deployment. It handles the subscription and notification traffic, and it rewrites the host portion of the URI before forwarding if the user is detected to be on a remote Cisco IM and Presence Service cluster within the same domain.

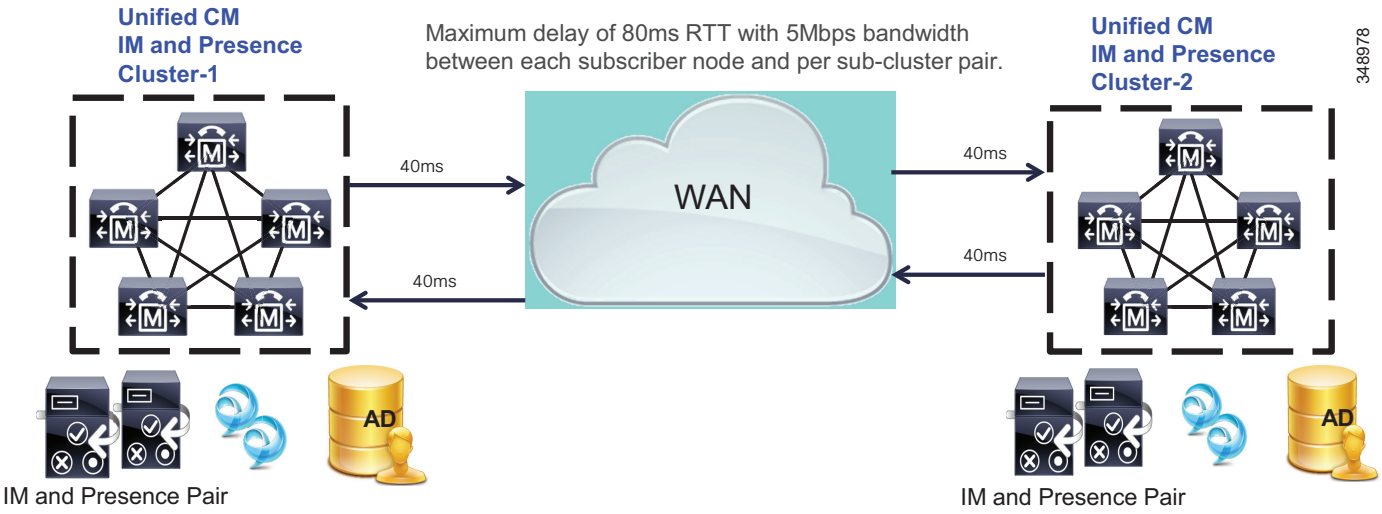
Clustering Over the WAN

A Cisco IM and Presence Service cluster can be deployed with one of the nodes of a subcluster deployed across the Wide Area Network (WAN). This allows for geographic redundancy of a subcluster and high availability for the users between the nodes across the sites. The following guidelines must be used when planning for a Cisco IM and Presence Service deployment with clustering over the WAN.

Geographic Data Center Redundancy and Remote Failover

A Cisco IM and Presence Service cluster can be deployed between two sites with a single subcluster topology, where one server of the subcluster is in one geographic site and the other server of the subcluster is in another site. This deployment must have a minimum bandwidth of 5 Mbps, a maximum latency of 80 ms round-trip time (RTT), and TCP method event routing (see Figure 20-12).

Figure 20-12 Intra-Cluster Bandwidth and Delay



High Availability and Scale

Cisco IM and Presence Service high availability allows for users on one node within a subcluster to automatically fail-over to the other node within the subcluster. With a Cisco IM and Presence Service subcluster containing a maximum of two nodes, remote failover is essentially between two sites, one site for each node. A scalable highly available capacity for a Cisco IM and Presence Service cluster is up to three subclusters; therefore, a scalable highly available remote failover topology would consist of the following two sites:

- Site A: Subcluster 1 node A, subcluster 2 node A, and subcluster 3 node A
- Site B: Subcluster 1 node B, subcluster 2 node B, and subcluster 3 node B

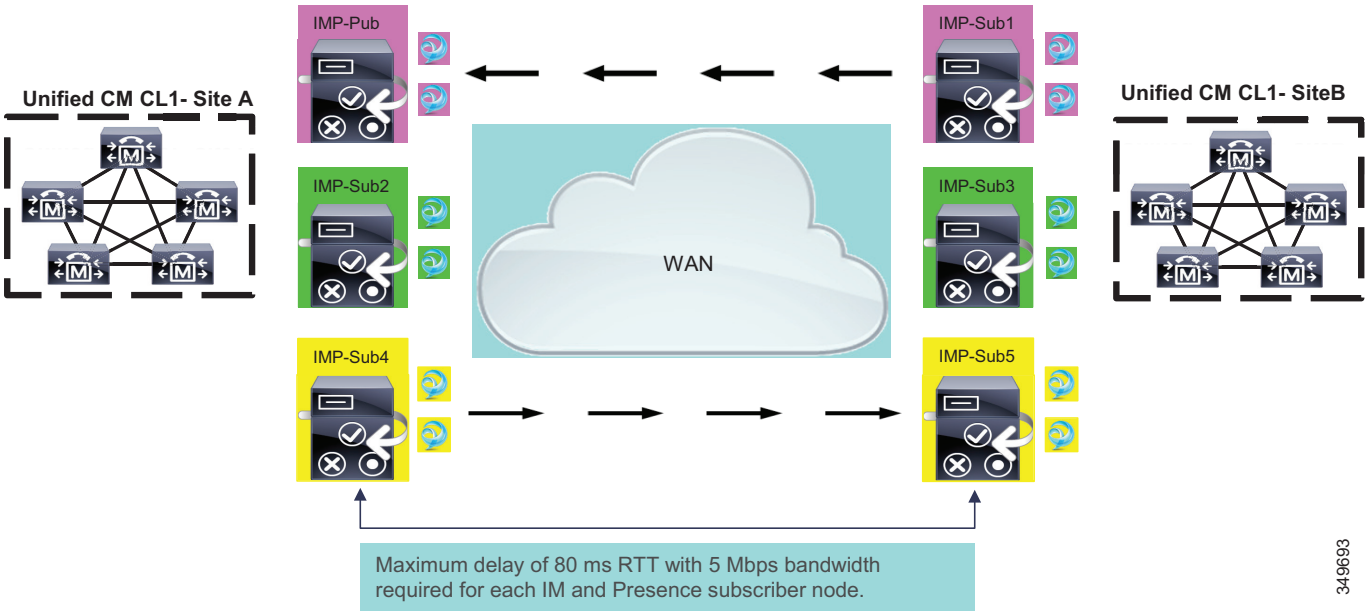
This deployment must have a minimum bandwidth of 5 Mbps per subcluster, a maximum latency of 80 ms round-trip time (RTT), and TCP method event routing. Each new subcluster added to the deployment requires an additional 5 Mbps of dedicated bandwidth to handle the database and state replication.

In the split subcluster model, the subcluster pairs are split across the WAN and deployed at two separate locations. The split subcluster model supports a maximum of 6 nodes deployed at 6 different locations, assuming the bandwidth and delay requirements are followed, with a maximum round-trip time (RTT) of 80 ms and 5 Mbps of bandwidth for each of the 6 nodes.

Split subcluster deployments have the following requirements:

- The Unified CM publisher and IM and Presence publisher must reside on the same side of the WAN, otherwise issues can arise with upgrades, especially with refresh (dual) upgrades.
- All IM and Presence nodes must have a minimum bandwidth of 5 Mbps, and an additional 5 Mbps for the cluster for database synchronization.
- Every IM and Presence node must also reside within and not exceed 80 ms RTT delay.
- All users must be distributed evenly across all split cluster nodes (see Figure 20-13). For example, assuming a subcluster supports 15,000 users, then each node of the split subcluster would support 7,500 on the 15k-user VM configuration template.

Figure 20-13 IM and Presence Subcluster Split Across the WAN



Local Failover

A Cisco IM and Presence Service cluster deployment between two sites may also contain a subcluster topology per site (single node or dual node for high availability), where one subcluster is in one geographic site and the other subcluster is in another geographic site. This topology allows for the users to remain at their local site (highly available or not) without the requirement or need to fail-over to a different site or location. This deployment must have a minimum bandwidth of 5 Mbps dedicated bandwidth between each subcluster in the respective sites, a maximum latency of 80 ms round-trip time (RTT), and TCP method event routing.

Bandwidth and Latency Considerations

With a Cisco IM and Presence Service cluster that has a topology of nodes split across a WAN, the number of contacts within a user's client can impact the bandwidth needs and criteria for the deployment. The traffic generated within and between Cisco IM and Presence Service clusters is directly proportional to the characteristics of the presence user profile, and thus the amount of bandwidth required for deployment. Cisco recommends 25% or fewer remote contacts for a client in environments where the bandwidth is low (10 Mbps or less), and at all times the maximum round-trip latency must be 80 ms or less.

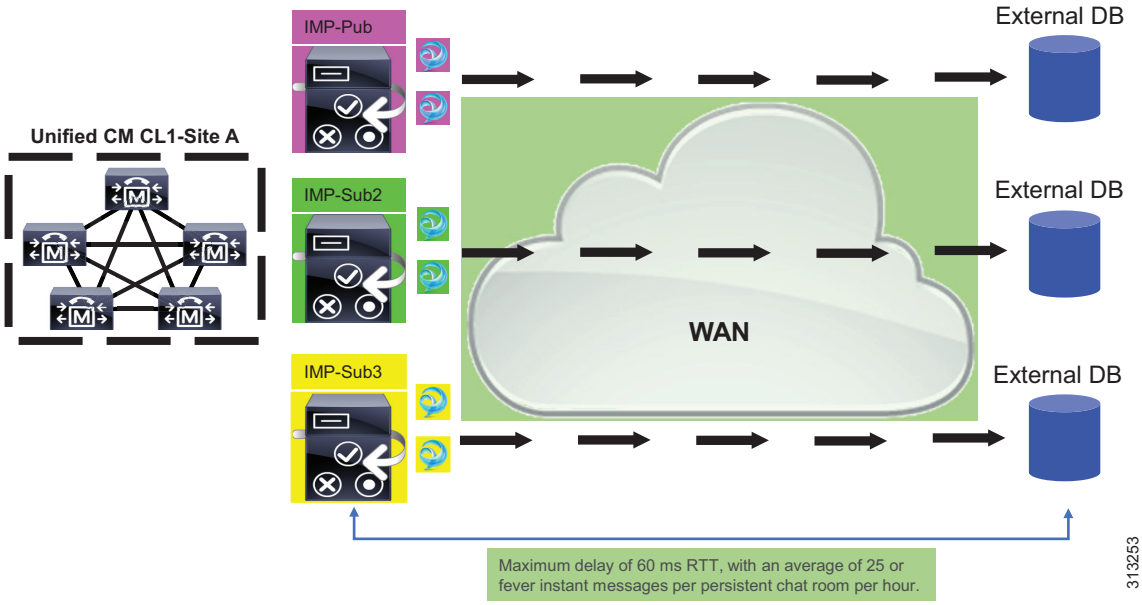
Persistent Chat and Compliance Logging Considerations

If Cisco IM and Presence Service is enabled for persistent chat, message archiving, or compliance logging, and if a subcluster is split across a WAN, then Cisco recommends placing the external database server(s) on the same side of the WAN as the Cisco IM and Presence Service subscriber node(s) with which it is associated.

With the ability to support multiple database instances on a single server and the recommendation for an external database server to reside on the same side of the WAN as its associated IM and Presence node. if a Cisco IM and Presence Service cluster is split across a WAN, then Cisco recommends deploying two external database servers as the best practice.

Note It is not a requirement for an external database server and its associated IM and Presence node(s) to be located in the same data center. However, the maximum supported latency between the external database server and the IM and Presence subscriber node(s) must not be greater than 60 ms round trip time (30 ms in each direction), and the minimum bandwidth allocation must align with clustering over the WAN requirements for the Cisco IM and Presence Service. Also, the average number of instant message per persistent chat room should not exceed 25 per hour. (See Figure 20-14.)

Figure 20-14 Persistent Chat with External Database Across the WAN



Note Cisco recommends using Oracle Database 12c for deploying an external database across the WAN and/or for high availability.

Centralized IM and Presence Deployments

Centralized IM and Presence can provide presence services to multiple remote Cisco Unified CM voice and video clusters. In a centralized deployment, the IM and Presence Service manages all presence related services for all the users across the remote Unified CM clusters, and each remote Unified CM cluster manages its own users' voice and video needs.

The centralized IM and Presence model provides an option for large enterprises to distribute the Unified CM clusters at multiple locations without the need to deploy an IM and Presence cluster at each of those locations. Hence, only a single centralized IM and Presence cluster is required for presence service for multiple remote Unified CM clusters.

Deployments with many locations but very few users at each location can benefit greatly from this model. For example, a hospital might have 50 locations with 100 users at each location that require voice, video, and presence service. It would not be feasible to deploy 50 clusters of Unified CM and 50 clusters of IM and Presence. Instead, a more efficient and simpler approach would be to use a single centralized IM and Presence cluster serving the 50 remote Unified CM clusters, thus greatly reducing cost with the reduction of virtual servers.

The centralized IM and Presence cluster must be deployed with the 15k/25k-user IM and Presence VM template. The Unified CM publisher for the centralized cluster must be deployed using the 10k-user Unified CM VM template. (See Figure 20-15.)

Note The centralized deployment design must be reviewed and approved via the Cisco Megacluster review process when more than 40,000 clients and/or devices are deployed.

For more details on configuring a centralized IM and Presence deployment, refer to the latest version of the guide on *Configuration and Administration of IM and Presence Service on Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Figure 20-15 Centralized IM and Presence Deployment

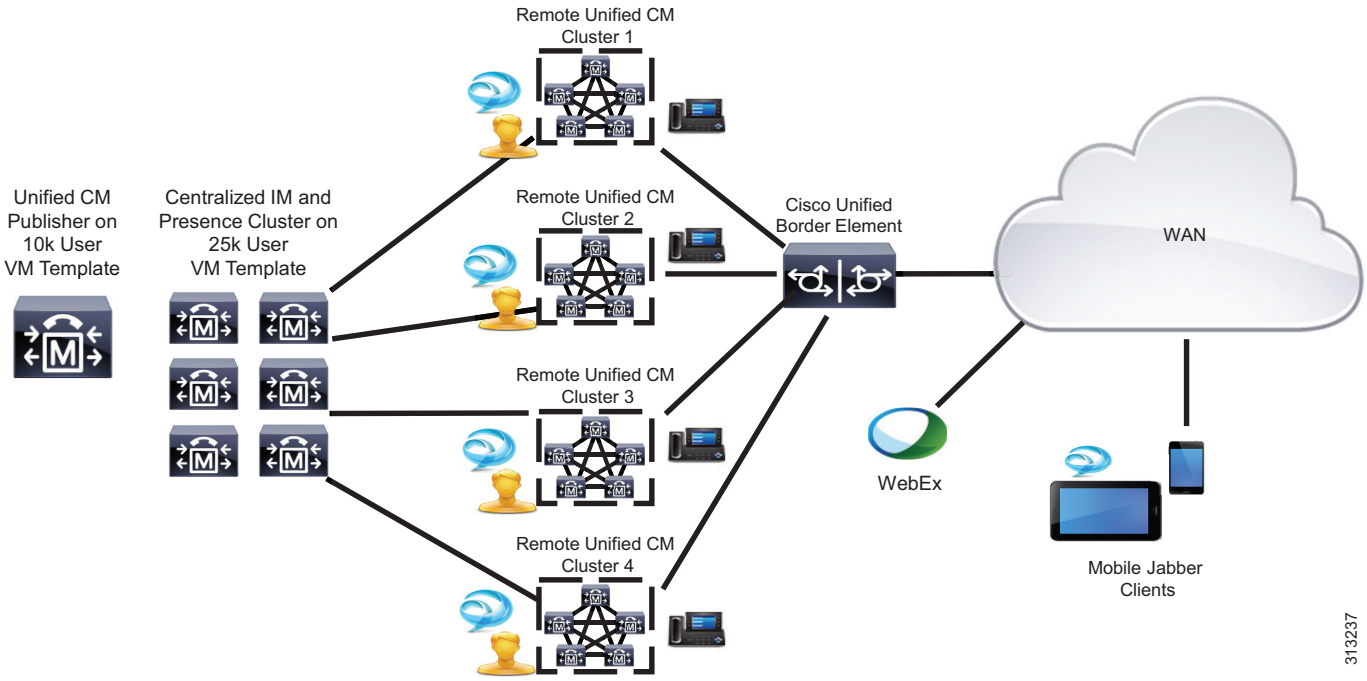


Figure 20-16 illustrates the following login flow steps for centralized IM and Presence when single sign-on (SSO) is not used:

- Steps 1 to 2: Query DNS to get the SRV record.
- Steps 3 to 4: Query UDS to get the home Unified CM cluster.
- Steps 5 to 8: Get the Access Token and Refresh Token from the Unified CM cluster through LDAP authentication.
- Step 9: Read **UC Service Profile -> IM& Presence Profile** and get the IM and Presence node information.
- Step 10: The Jabber client registers to the IM and Presence cluster using the same Access Token through SOAP and XMPP interfaces.

- Step 11: The IP Multimedia Subsystem (IMS) API invokes the AuthZ service to validate the token, and the response is sent back to the Jabber client.

Note The service parameter **Enable User for Unified CM IM and Presence** must be disabled (unchecked) on the remote Cisco Unified CM clusters.

Figure 20-16 Centralized IM and Presence Login Flow without Single Sign-On (SSO)

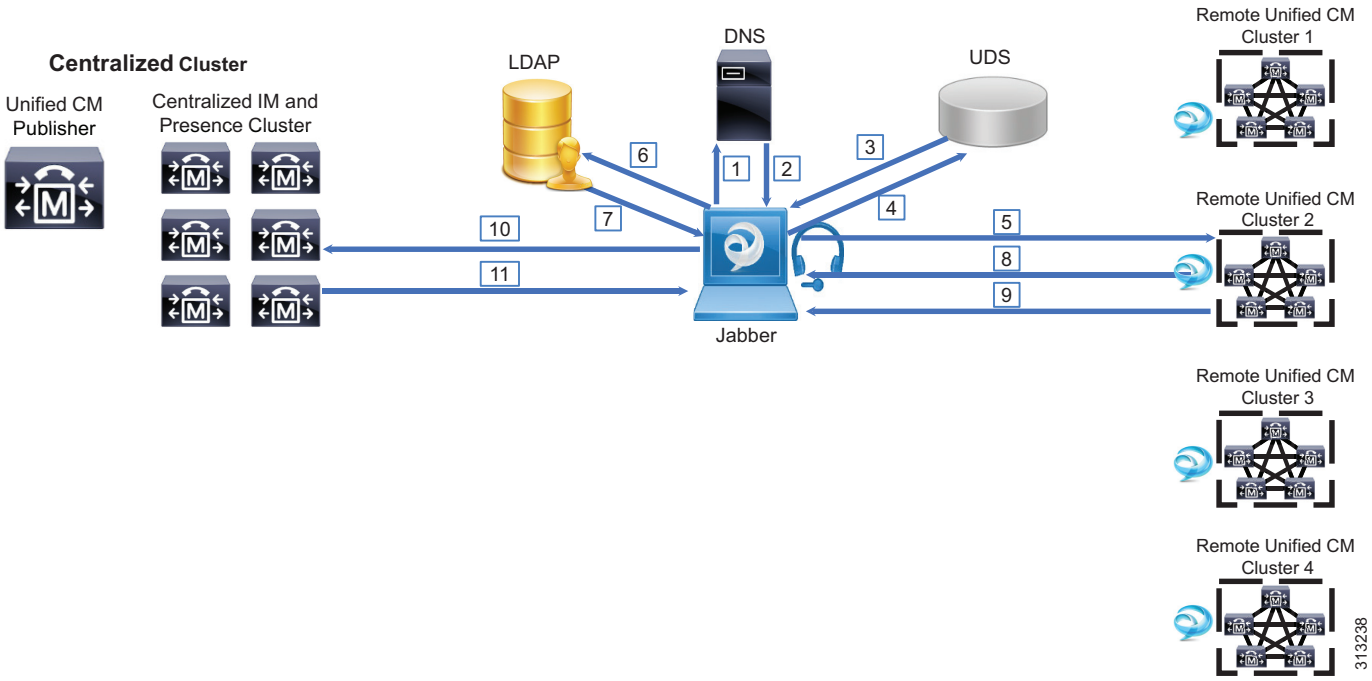


Figure 20-17 illustrates the following login flow steps for centralized IM and Presence when single sign-on (SSO) is used:

- Steps 1 to 2: Query DNS to get the SRV record.
- Steps 3 to 4: Query UDS to get the home Unified CM cluster.
- Steps 5 to 8: Get the Access Token and Refresh Token from the Unified CM cluster through Identity Provider (IdP) authentication.
- Step 9: Read **UC Service Profile -> IM& Presence Profile** and get the IM and Presence node information.
- Step 10: The Jabber client registers to the IM and Presence cluster using the same Access Token through SOAP and XMPP interfaces.
- Step 11: The IP Multimedia Subsystem (IMS) API invokes the AuthZ service to validate the token, and the response is sent back to the Jabber client.

Note The service parameter **Enable User for Unified CM IM and Presence** must be disabled (unchecked) on the remote Cisco Unified CM clusters.

Figure 20-17 Centralized IM and Presence Login Flow with Single Sign-On (SSO)

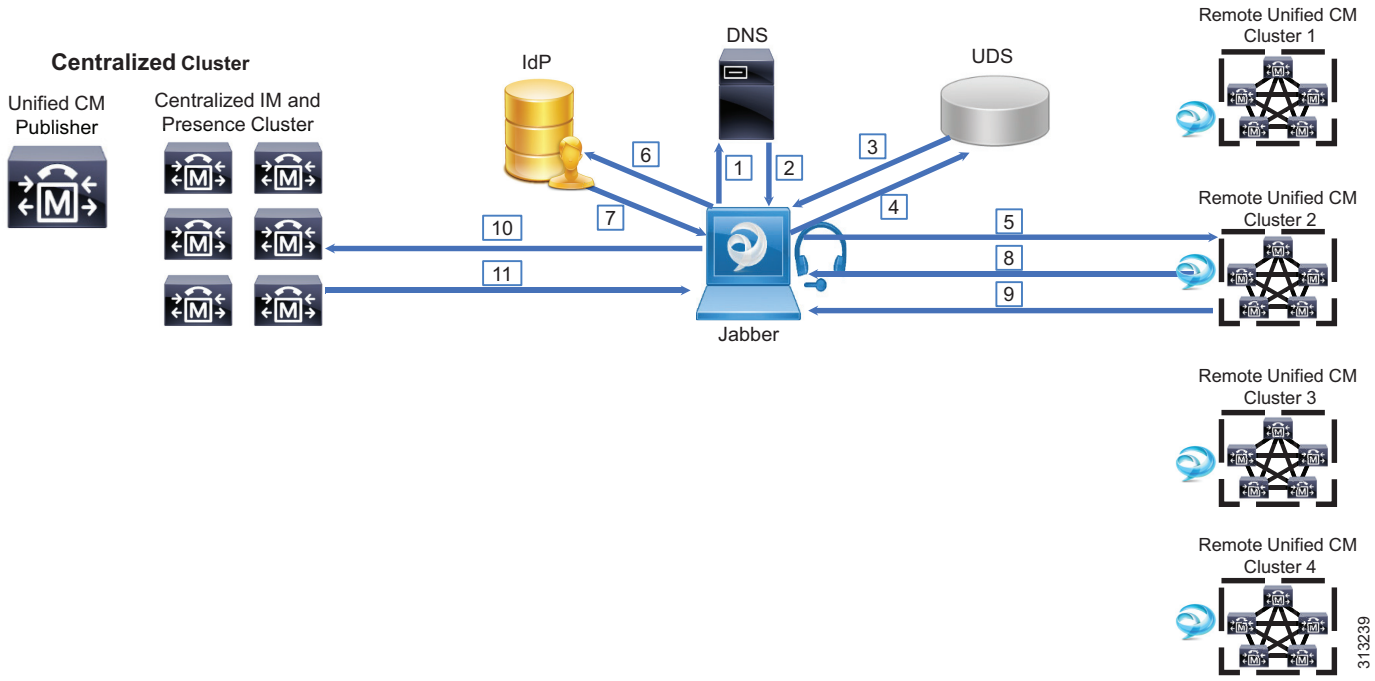
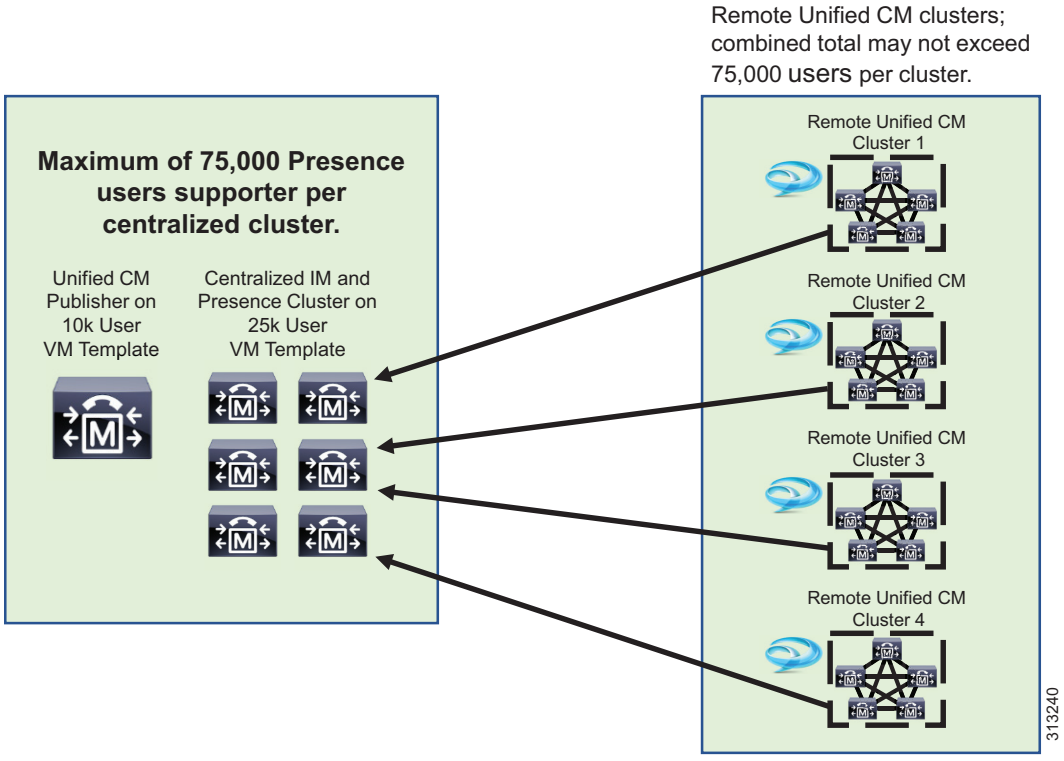


Figure 20-18 Maximum Number of Supported Users for Centralized IM and Presence



Federated Deployment

Cisco IM and Presence Service allows for business-to-business communications by enabling inter-domain federation, which provides the ability to share presence and instant messaging communications between different domains. Inter-domain federation requires explicit DNS domains to be configured, as well as a security appliance (Cisco Adaptive Security Appliance) in the DMZ to terminate federated connections with the enterprise.

Multi-Domain Support

The IM and Presence Service provides the ability to configure more than a single domain for federation. The domains are automatically discovered by the system when using DirectoryURI, or the administrator can add the domains manually. When a federated deployment involves multiple domains, then DNS SRV records need to be published for each email domain. Each DNS SRV record should resolve to an identical set of results where XMPP federation is a list of all XMPP federation nodes and SIP federation is the Public FQDN of the Routing IM & Presence node.

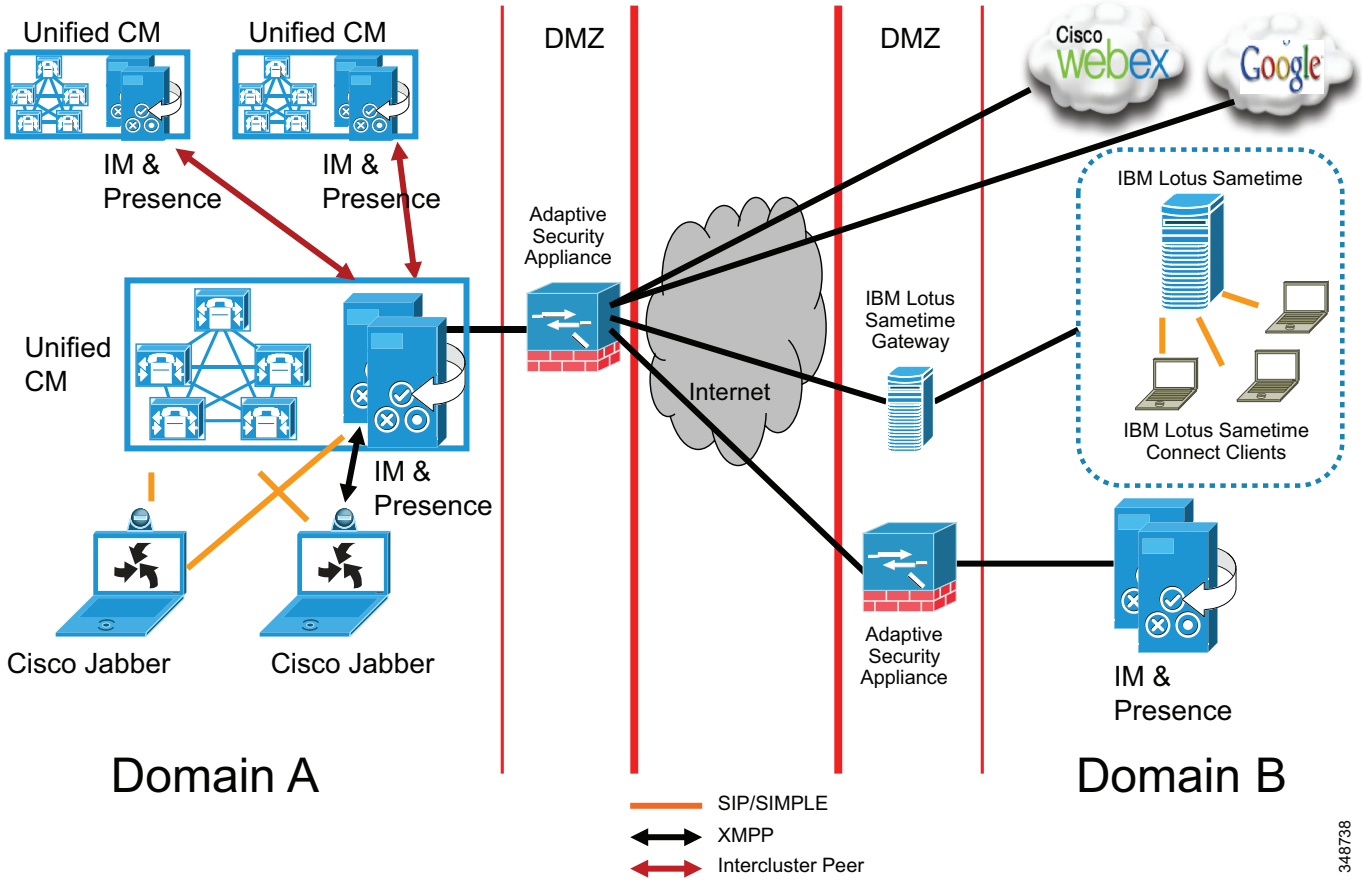
Federation with multiple email domains also requires regeneration of the security certificates cup-xmpp (certificate presented to XMPP clients) and cup-xmpp-s2s (certificate presented to federated systems). For both of these certificates, all the domains must be included as Subject Alt Name (SAN) entries. The manual administrative configuration gives the administrator the option to pre-populate the domains so that it is not necessary to regenerate the certificates every time a new domain automatically gets discovered.

If all the federated domains are within the same trust boundary, where a deployment has all components within a single datacenter, then the use of the Adaptive Security Appliance is not required. For information on inter-domain federation, refer to the latest version of *Interdomain Federation for IM and Presence Service on Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Figure 20-19 shows the basic inter-domain federation deployment between two different domains, indicated by Domain A and Domain B. The Adaptive Security Appliance (ASA) in the DMZ is used as a demarcation into the enterprise. XMPP traffic is passed through, whereas SIP traffic is inspected. All federated incoming traffic is routed through the Cisco IM and Presence Service that is enabled as a federation node, and is routed internally to the appropriate server in the cluster where the user resides. For intercluster deployments, intercluster peers propagate the traffic to the appropriate home cluster within the domain. All federated outgoing traffic is directed outward through any node in the IM and Presence Service cluster that has XMPP federation enabled. Multiple nodes can be enabled as federation nodes within large enterprise deployments, where each request is routed based on a round-robin implementation of the data returned from the DNS SRV lookup.

Figure 20-19 IM and Presence Service XMPP Federation (Inter-Domain)



Cisco IM and Presence Service also provides configuration through SIP to allow for inter-domain federation with Microsoft and AOL, as depicted in Figure 20-20. Cisco IM and Presence Service inter-domain federation with Microsoft Lync Server provides basic presence (available, away, busy, offline) and point-to-point instant messaging. Rich presence capability (On the Phone, In a Meeting, On Vacation, and so forth), as well as advanced instant messaging features, are not supported. Cisco IM and Presence Service inter-domain federation with AOL allows federation with users of AOL public communities (aim.com, aol.com), with users of domains hosted by AOL, and with users of a far-end enterprise that federates with AOL (that is, AOL is being used as a clearing house).

Note A SIP federation (inter-domain to AOL) on Cisco IM and Presence Service must be configured for each domain of the AOL network, which can consist of both hosted networks and public communities. Each unique hosted domain must be configured; however, only a single aol.com public community needs to be configured because the AOL network allows a user to be addressed as user@aol.com or user@aim.com

Figure 20-20 IM and Presence Service SIP Federation (Inter-Domain)

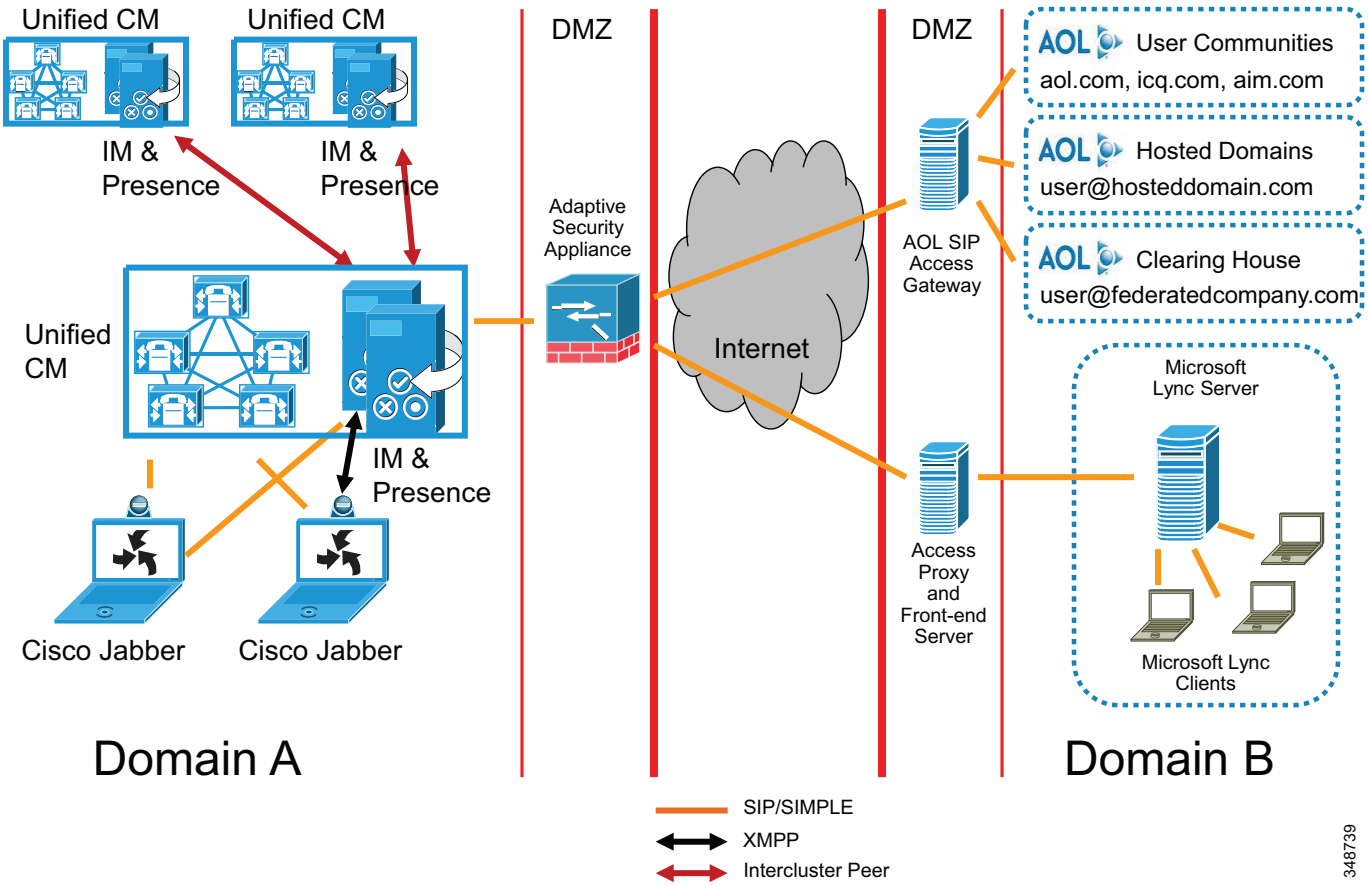


Table 20-2 lists the state mappings between Cisco IM and Presence Service and Microsoft Lync Server.

Table 20-2 Mapping of Presence States

Cisco Status	Cisco Color	Status to Microsoft Lync Server	Status to AOL
Do not disturb	Red	Busy	Away
Busy	Yellow	Busy	Away
On the phone	Yellow	Busy	Away
In a meeting	Yellow	Busy	Away
Idle on all clients	Yellow	Away	Away
Available	Green	Available	Available
Unavailable/offline	Grey	Offline	Offline

Note Cisco IM and Presence Service must publish a DNS SRV record (SIP, XMPP, and each text conferencing node) for the domain to allow for other domains to discover the Cisco IM and Presence Services through DNS SRV. With a Microsoft Lync Server deployment, this is required because Cisco IM and Presence Service is configured as a Public IM Provider on the Access Edge server. If the Cisco IM and Presence Service cannot discover the Microsoft domain using DNS SRV, you must configure a static route on Cisco IM and Presence Service for the external domain.

The Cisco IM and Presence Service SIP federation deployment can be configured with redundancy using a load balancer between the Adaptive Security Appliance and the Cisco IM and Presence Service, or redundancy can also be achieved with a redundant Adaptive Security Appliance configuration. For XMPP federation, redundancy can be achieved using DNS SRV records.

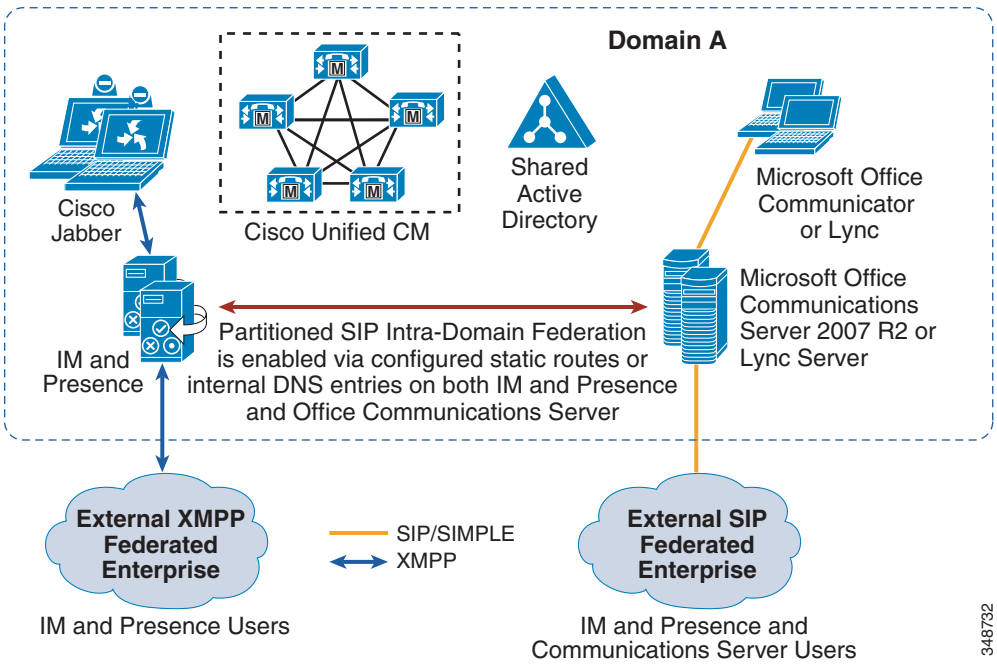
For additional configuration and deployment considerations regarding a federated deployment, refer to the latest version of *Interdomain Federation for IM and Presence Service on Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

An intra-domain partitioned federated deployment, shown in Figure 20-21, is a secondary option that allows for Cisco IM and Presence Service and Microsoft Lync Server to federate presence and instant messaging within the same presence domain. The users are partitioned across both deployments, within the single presence domain, and are licensed either on Cisco IM and Presence Service or on the Microsoft Lync Server.

Note The user cannot be licensed on both the Cisco and Microsoft platforms at the same time.

Figure 20-21 Cisco IM and Presence Service Intra-Domain Federation



The partitioned intra-domain federation between the Cisco and Microsoft platforms is based on the SIP/SIMPLE protocol and allows for basic presence and instant messaging exchange, as supported with the Cisco IM and Presence Service inter-domain federation support for Microsoft. Rich presence and group chat functionality are not supported with the partitioned intra-domain presence federation.

Inter-domain federation and partitioned intra-domain federation can be supported simultaneously with the following qualifications:

- XMPP federation may be enabled on the Cisco IM and Presence Service deployment but is available only to Cisco IM and Presence Service licensed users.
- SIP federation may be enabled either on Cisco IM and Presence Service or on Microsoft Office Communications Server 2007 R2 or Lync Server; however, for SIP Federation to be available to both Cisco and Microsoft users, it must be enabled on Microsoft Office Communications Server 2007 R2 or Lync Server.
- If SIP/SIMPLE inter-domain federation with Microsoft Lync or Office Communications Server is required in parallel with the partitioned intra-domain federation, then the Microsoft Office Communications Server or Lync Server can be configured to manage that external federation. Cisco IM and Presence Service administration must be configured with static routes to the Microsoft environment for the external domain. Alternatively, Cisco IM and Presence Service could manage the SIP federation, while Microsoft Lync or Office Communications Server could manage the XMPP federation.

On-Premises Cisco IM and Presence Service SAML SSO for Jabber

Note The supported end point for SAML SSO deployment is Cisco Jabber.

The Security Assertion Markup Language Single Sign-On (SAML SSO) feature enhances the end user experience by avoiding the need to log in multiple times to multiple applications within the collaboration solution.

SAML SSO provides a secure mechanism to use credentials and relevant information of the end user across multiple Unified Communications applications such as Unified CM, Cisco Unity Connection, IM and Presence, Jabber clients, and so on.

For the SAML SSO feature to work correctly, ensure that the network architecture scales to meet the number of users for each cluster, assuming that each user may have as many as five or more services that require authentication and a minimum of two devices associated to each user. For a deployment across multiple Unified Communications applications, all SAML requests must authenticate with the IdP for Cisco Jabber clients to log in successfully.

Note SSO is supported by Unified Communications services with SAML and OAuth only.

Cisco Jabber with SAML SSO does impact performance during logins, and the current maximum login rate for 5,000 users is within half an hour. This is assuming that you have distributed devices and users evenly across all nodes and that Cisco Jabber is in softphone mode.

Cisco Jabber is the only supported client and/or endpoint for IM and Presence deployments that supports SAML SSO.

For sizing information and examples, see the section on SAML SSO Cisco Jabber Client, page 25-20, in the chapter on Collaboration Solution Sizing Guidance, page 25-1.

On-Premises Cisco IM and Presence Service Enterprise Instant Messaging

Cisco IM and Presence Service incorporates the supported enterprise instant messaging features of the Extensible Communications Platform (XCP), while allowing for some modifications to enhance support for multi-device user experience. Cisco IM and Presence Service changes the XCP instant messaging routing architecture to allow for initial instant messages to be routed to all of the user's non-negative priority logged-in devices, rather than routing to the highest priority device as is done with existing XCP installations. Backward compatibility support for point-to-point instant messaging between Cisco IM and Presence Service SIP clients and XMPP clients is provided by IM internal gateway functionality.

The IM and Presence Service supports IM exchange in both ad hoc chat rooms and persistent chat rooms. By default, the Text Conference (TC) component on the IM and Presence Service is set up and configured to handle IM exchange in ad hoc chat rooms.

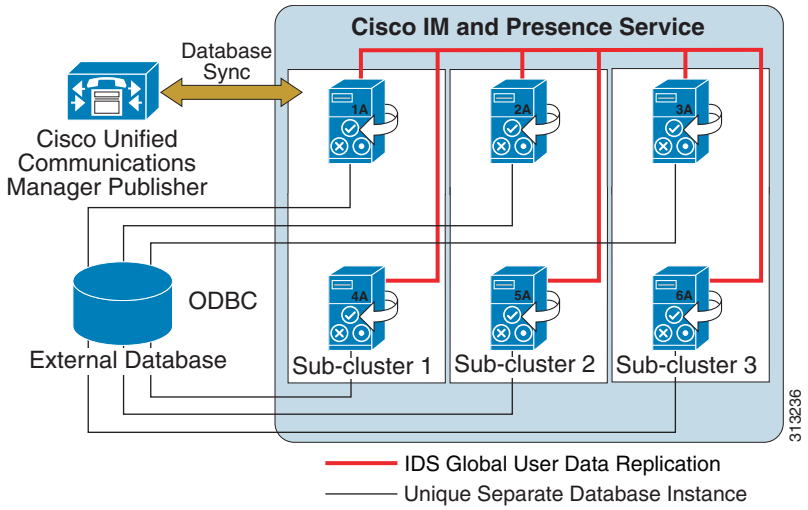
Ad hoc chat rooms are IM sessions that remain in existence only as long as one person is still connected to the chat room, and they are deleted from the system when the last user leaves the room. Records of the IM conversation are not maintained permanently.

Persistent chat rooms are group chat sessions that remain in existence even when all users have left the room, and they do not terminate like ad hoc group chat sessions. The intent is that users can return to persistent chat rooms over time to collaborate and share knowledge of a specific topic, search through archives of what was said on that topic, and then participate in the discussion of that topic in real-time.

For persistent chat you must have 1:1 mapping of the external database instance for each of the nodes in the cluster. The size of the database should be taken into consideration. Archiving messages is optional, and it increases the traffic on the node to which the external database instance its attached. In large deployments, disk space is a concern because it can be filled very quickly, so you must ensure that your database is large enough to handle the volume of information being logged.

Cisco IM and Presence Service uses the basic interfaces of the external database and does not provide any administration, interface hooks, or configuration of the database. Cisco requires a separate database instance for each server in the cluster when Cisco IM and Presence Service is deployed with persistent group chat. (See Figure 20-22.) The database instances can share the same hardware but are not required to do so.

Figure 20-22 Cisco IM and Presence Service Persistent Chat

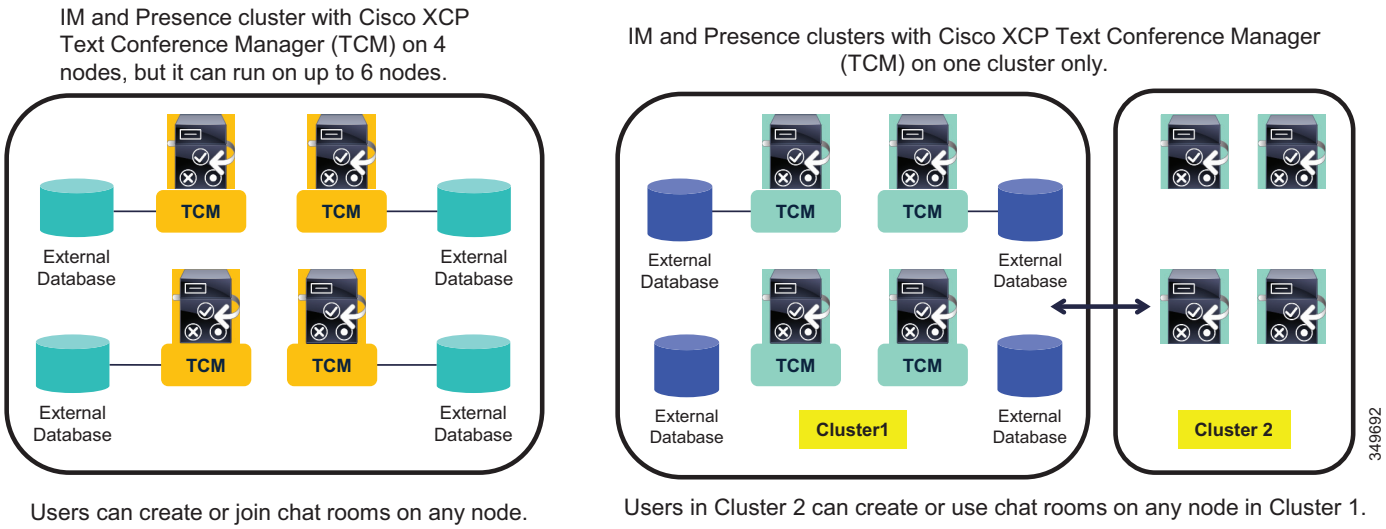


If persistent chat is enabled, an external database must be associated with the Cisco XCP Text Conference Manager service, and the database must be active and reachable otherwise the Text Conference Manager service will not start. If the connection with the external database fails after the Text Conference Manager service has started, the Text Conference Manager service will remain active and functional; however, messages will no longer be written to the database, and new persistent chat rooms cannot be created until the connection recovers.

Deployment Considerations for Persistent Chat

- Persistent chat can be deployed on one or more nodes in a cluster (see Figure 20-23).
- Each node that supports persistent chat must be assigned to a dedicated database instance.
- An external database server may support multiple database instances.
- Persistent chat is a cluster-wide setting.
- At least one node in the cluster must be assigned to an external database.
- The Cisco XCP Text Conference Manager service will not run on a node that is not assigned to an external database.
- Pure instant (ad hoc) conferencing does not require any external database.

Figure 20-23 Persistent Chat Deployment



Chat Room Limits for IM and Presence Service

The IM and Presence Service supports point-to-point file transfer between XMPP clients. Table 20-3 lists the chat room limits for the IM and Presence Service

Table 20-3 Chat Room Limits for IM and Presence Service

Number of	Maximum
Persistent chat rooms per node	1,500 rooms
Total rooms per node (ad hoc and persistent)	16,500 rooms
Occupants per room	1,000 occupants
Messages retrieved from the archive. This is the maximum number of messages that are returned when a user queries the room history.	100 messages
Messages in chat history displayed by default. This is the number of messages that are displayed when a user joins a chat room.	15 messages

Text conferencing, sometimes referred to as multi-user chat, is defined as ad-hoc group chat and persistent group chat and is supported as part of the XCP feature set. In addition, offline instant messaging (storing instant messages for users who are currently offline) is also supported as part of the XCP feature set. Cisco IM and Presence Service handles storage for each of these instant messaging features in different locations. Offline instant messaging is stored locally in the Cisco IM and Presence Service IDS database.

Ad-hoc group chat is stored locally in memory on the Cisco IM and Presence Service. Persistent group chat requires an external database to store chat rooms and conversations. The external databases supported are PostgreSQL (see https://www.postgresql.org/), Microsoft SQL, and Oracle (see https://www.oracle.com).

Note Cisco does not provide any database best practices or any data extraction tools. Those tasks and tools are expected to be provided by a database administrator.

Note When Oracle is used as the external database, tablespace information must be configured.

Managed File Transfer

Managed file transfer (MFT) allows an IM and Presence Service client such as Cisco Jabber to transfer files to other users, ad hoc group chat rooms, and persistent chat rooms. The files are stored in a repository on an external file server, and the transaction is logged to an external database.

This configuration is specific to file transfers and has no impact on the message archiver feature for regulatory compliance.

Note There is no high availability solution for MFT or persistent chat because there is only one connection from the IM and Presence node to each external third-party database.

Software

- Cisco IM and Presence Service, Release 10.5(2) or later
- PostgreSQL or Microsoft SQL
- Oracle, versions 9i, 10g, or 11g
- File server versions Centos 6.5 or greater

Note Oracle Data Guard may be used as an external database, but it has not been tested by Cisco.

Note If you are required to have an encrypted connection to the external database, you must use Oracle 11g. No other supported database version allows encrypted connections.

You can install the database on either a Linux or a Windows operating system. See the PostgreSQL, Microsoft SQL, and Oracle documentation for details on the supported operating systems and platform requirements.

IPv4 and IPv6 are supported, as is dual-stack mode.

Transfer Process

The flow for transferring a file to a single recipient involves the following steps:

1. The sender's client uploads the file via HTTP, and the server responds with a URI for the file.
2. The file is stored in the repository on the file server.
3. An entry is written to the external database log table to record the upload.
4. The sender’s client sends an IM to the recipient; the IM includes the URI of the file.
5. The recipient’s client requests the file via HTTP.

6. The file is read (retrieved) from the repository.
7. The download request is recorded in the log table.
8. The file is downloaded to the recipient.

The flow for transferring a file to a group chat or persistent chat room is similar, except that the sender sends the IM to the chat room, and each chat room participant sends a separate request to download the file.

Managed File Transfer on IM and Presence Service

When you enable managed file transfer on an IM and Presence Service node, consider the following information:

- If you deploy any combination of the persistent group chat, message archiver, or managed file transfer features on an IM and Presence Service node, you can assign the same physical external database installation and file server to all of these features. However, you should consider the potential IM traffic and file transfers (size and number) when you determine the server capacity.
- The node public key is invalidated if the node's assignment is removed. If the node is reassigned at a later date, a new node public key is automatically regenerated. The external file server will also need to be reconfigured.
- The Cisco XCP File Transfer Manager service must be active on each node where managed file transfer is required.

Managed File Transfer Capacities

All Jabber users have the option to transfer files at will, which can potentially impact the system based on the file transfer usage. Refer to Table 20-4 to help calculate your capacity needs.

The values in Table 20-4 are based strictly on a transfered file size of 500 kilobytes (KB). These value may be adjusted to calculate different capacities; for example, any one of the following scenarios is equivalent to 1,500 transfers per hour:

- 1,500 users, each downloading or uploading a single 500 KB file in an hour
- 3,000 users, each transferring a 250 KB file per hour
- 750 Jabber users sending a single 500 KB file to 750 other Jabber users

The maximum number of supported transfers is 12,000 with a 500 KB file.

Table 20-4 Jabber File Transfer Capacities

Usage Level	Transfers per Hour	CPU % Total	CPU % AFT	AFT_LOG Table	AFT_LOG Size	JM Table Additional Size
Low usage	1,500	35%	25%	3,000	0.6 MB	1.5 MB
Medium usage	4,500	50%	40%	9,000	2.8 MB	4.5 MB
Maximum usage	12,000	65%+	55%	24,000	7.8 MB	12.0 MB

The File Transfer window has the following File Transfer Type controls:

- Disabled — File transfer is disabled for the cluster.
- Peer-to-Peer — One-to-one file transfers are allowed, but files are not archived or stored on a server. Group chat file transfer is not supported.

Managed File Transfer and Persistent Group Chat both require an external database instance per IM and Presence node in an IM and Presence cluster.

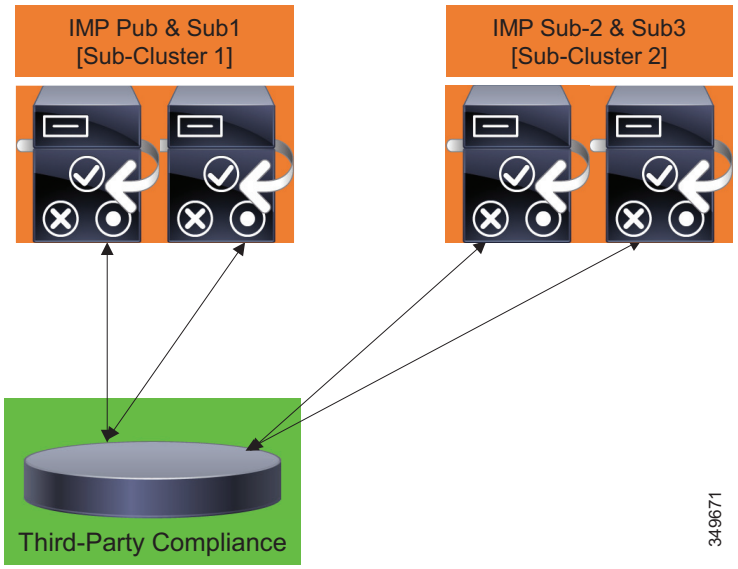
Note A node that has Managed File Transfer enabled should not be deployed in a cluster with a node that has Peer-to-Peer enabled. The recommended migration path is to configure the Peer-to-Peer nodes as Managed and Peer-to-Peer File Transfer nodes, and then change them to Managed File Transfer nodes.

On-Premises Cisco IM and Presence Service Message Archiving and Compliance

As part of the architecture, Cisco IM and Presence Service contains a Message Archiver component that allows for logging of point-to-point, text conferencing, federated, and intercluster messages into an external database as part of a non-blocking compliance. Cisco IM and Presence Service message archival requires an external database (PostgreSQL, Microsoft SQL, or Oracle) instance per cluster. The same database can be shared with multiple clusters; however, a large number of users in a intercluster peer deployment would require more database instances due to capacity demands and a high number of data inserts. Although one external database instance per IM and Presence cluster is supported, the minimum recommendation is to deploy one external database instance per sub-cluster pair.

A blocking third-party compliance solution, which not only allows logging of messages but also applies policy to message delivery and message content, is provided through a third-party compliance server solution, as illustrated in Figure 20-24. Cisco IM and Presence Service third-party compliance can be deployed with multiple compliance servers for each server in the cluster, multiple servers per compliance server, or some other combination. Use of a third-party compliance solution is mutually exclusive with using the Message Archiver feature.

Figure 20-24 On-Premises Cisco IM and Presence Service: Third-Party Compliance



All Cisco IM and Presence Service servers in the cluster are subject to compliance. Figure 20-25 shows a deployment with a compliance server for each server in the IM and Presence Service cluster; whereas Figure 20-26 shows a mapping of a single compliance server to multiple IM and Presence Service servers, or multiple compliance servers to a single IM and Presence Service server. The various deployment options allow for greater flexibility in compliance policy routing and cluster deployment.

Figure 20-25 Cisco IM and Presence Service Third-Party Compliance

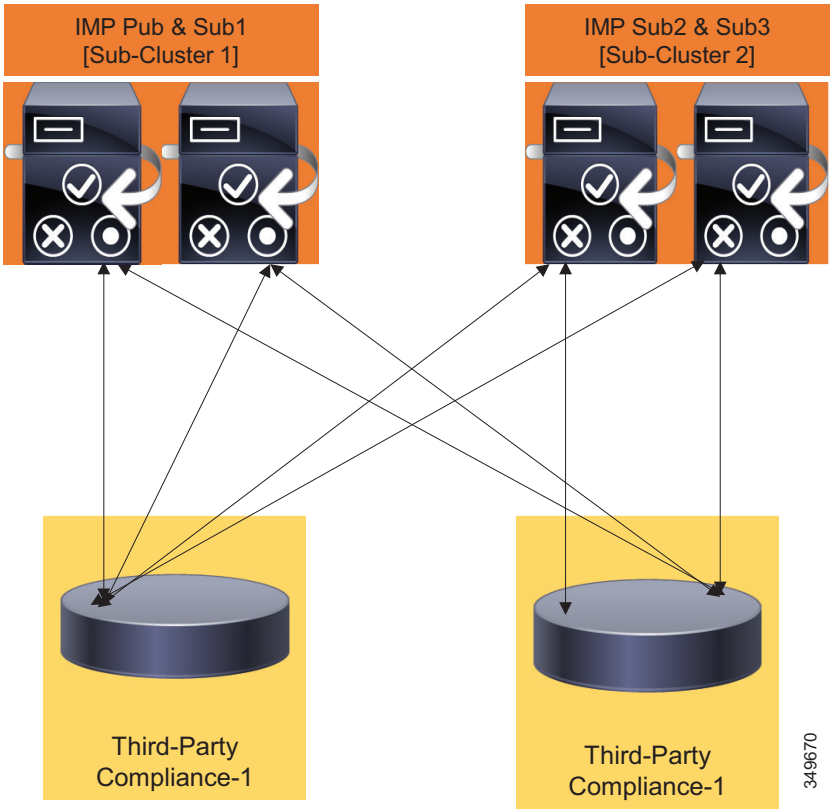
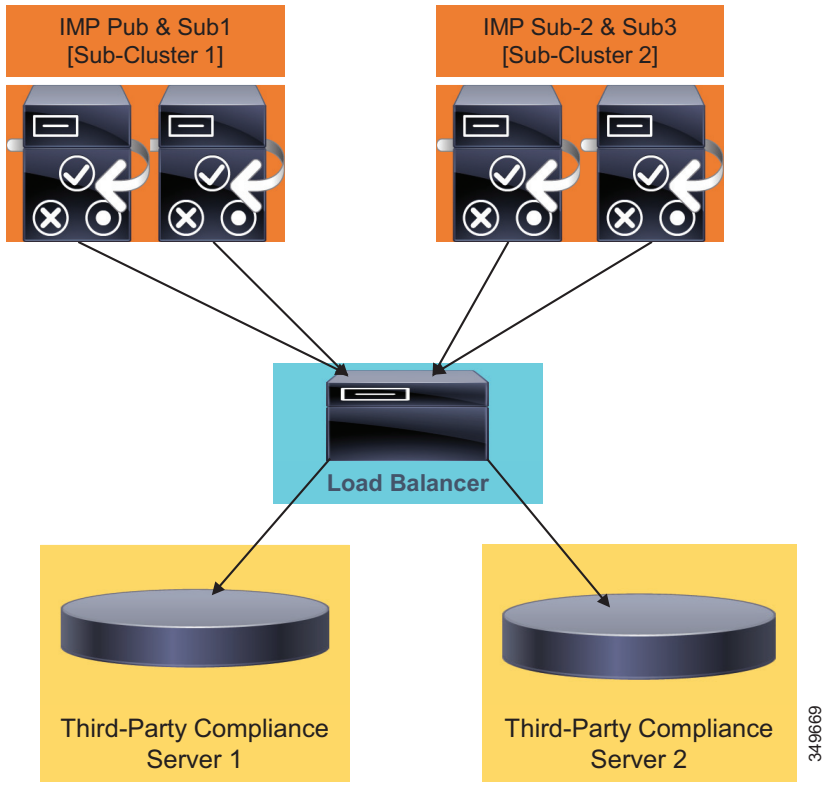


Figure 20-26 Full High Availability Cluster-Wide Compliance



Cluster-wide compliance allows for configuration of compliance profiles based on particular events, while allowing those events to be prioritized and routed to the appropriate compliance servers when there are overlapping events in the compliance profiles. Every compliance server must have a compliance profile assigned, and multiple compliance servers can share the same compliance profile.

Requirements for the IM and Presence Service external database differ depending on the feature(s) utilized. For example, enabling the Persistent Group Chat feature requires an external database for each IM and Presence sub-cluster pair, where every IM and Presence sub-cluster pair in the cluster points to a unique database instance but can share the same physical database installation.

The Message Archiver (compliance) feature requires at least one external database instance per IM and Presence cluster. The recommendation, however, is to have 1:1 mapping with two external database instances per sub-cluster pair and to define each IM and Presence node in the compliance profile assigned to the respective compliance server(s).

The Managed File Transfer feature requires one unique logical external database instance for each IM and Presence Service node in an IM and Presence Service cluster that has the Cisco XCP File Transfer Manager service activated.

Note If you deploy any combination of the persistent group chat, message archiver (compliance), and managed file transfer features on an IM and Presence Service node, you can assign the same external database to each feature.

Collaboration Client Message Logging Storage Requirements

The message archiving and Persistent Chat functionality use an external database to store messages offline. There are a number of factors to consider for the storage requirements of a deployment, such as the customer topology, how the database is tuned, and how messaging is used within the organization. The following calculations provide guidelines for these inputs to be used in estimating the raw database storage requirements of a deployment for external database storage.

Cisco IM and Presence Service supports both SIP and XMPP clients, and there are slightly different amounts of overhead per message based on the protocol. The overhead per message for message archiving could actually be larger or smaller depending on deployment, Jabber Identifier/UserID size, client type, and thread ID; therefore, an average overhead amount is used. For SIP-based messages the average overhead is 800 bytes and for XMPP messages the average overhead is 600 bytes.

The minimum storage requirements (in bytes) for message archiving per month for Cisco Jabber users can be calculated as follows:

(Number of users) * (Number of messages/hour) * (Number of busy hours/month) * (600 + (3 * Number of characters/message))

The message archiving requirements above must be doubled if **Enable Outbound Message Logging** is enabled on Cisco IM and Presence Service compliance configuration.

The minimum storage requirements (in bytes) for persistent chat per month for Cisco Jabber users can be calculated as follows:

(Number of users) * (Number of Persistent Chat messages/hour) * (Number of busy hours/month) * (700 + (3 * Number of characters/message))

Note Persistent Chat is supported only with XMPP clients and uses an average overhead of 700 bytes.

Table 20-5 provides an example spreadsheet to assist in determining the database space requirements. These calculations are provides as a very simplified calculation. This example does not attempt to provide any differentiation between database types or how storage may be used.

Table 20-5 Example Spreadsheet for Estimating Database Storage Requirements

	A	B	C	D	E
1	**Description**	**Message Archiver**	**Persistent Chat**	**Managed File Transfer**	**Total**
2	Feature Enabled	Yes	Yes	Yes	
3	Message Archiver Outbound Message Logging Enabled	No			
4	Number of users	2500	2500	2500	
5	Estimated number of messages per hour per user	15	15	2	
6	Number of busy hours per month	200	200	200	
7	Average number of characters per message	250	250	250	
8	XMPP message overhead bytes per message	600	700	600	
9	Database text message encoding factor	3	3	3	
10	Percent buffer allowance multiplier	150.00%	150.00%	150.00%	
11	**Computations**				
12	Number of messages per month	7,500,000	7,500,000	1,000,000	16,000,000
13	Formula for above calculation	IF(B2="Yes", IF(B3="Yes", 2,1)* ROUNDUP(B4*B5*B6,0),0)	IF(C2="Yes", ROUNDUP(C4*C5*C6, 0),0)	IF(D2="Yes", ROUNDUP(D4*D5*D6, 0),0)	
14	Estimated total GBs storage used per month	9.9	10.7	1.4	22.0
15	Formula for above calculation	ROUNDUP((B12* ((B7*B9)+B8))/ 1024000000,1)	ROUNDUP((C12* ((C7*C9)+C8)) / 1024000000,1)	ROUNDUP((D12* ((D7*D9)+D8)) / 1024000000,1)	
16	Estimated total database GBs to provision per month	14.9	16.1	2.1	33.1

These message archive and Persistent Chat numbers are the minimum storage requirements based on an average over time; therefore, a buffer multiplier of 1.5 (150%) should be used to account for very large UserIDs, larger than expected instant message lengths, and other factors that tend to increase the storage requirements. Table 20-6 lists some examples of storage requirements for Cisco Collaboration Clients.

Table 20-6 Examples of Cisco Collaboration Client Message Logging Storage Requirements

Profile	Number of Users	Number of Messages per Hours	Number of Busy Hours per Month	Average Size of Message	Message Archive Storage Requirement	Persistent Chat Storage Requirement
Light	1,500	10	200	100	2.7 GB	3.0 GB
Medium	2,500	15	200	250	9.9 GB	10.7 GB
High	2,500	25	200	500	25.7 GB	26.9 GB

On-Premises Cisco IM and Presence Service Calendar Integration

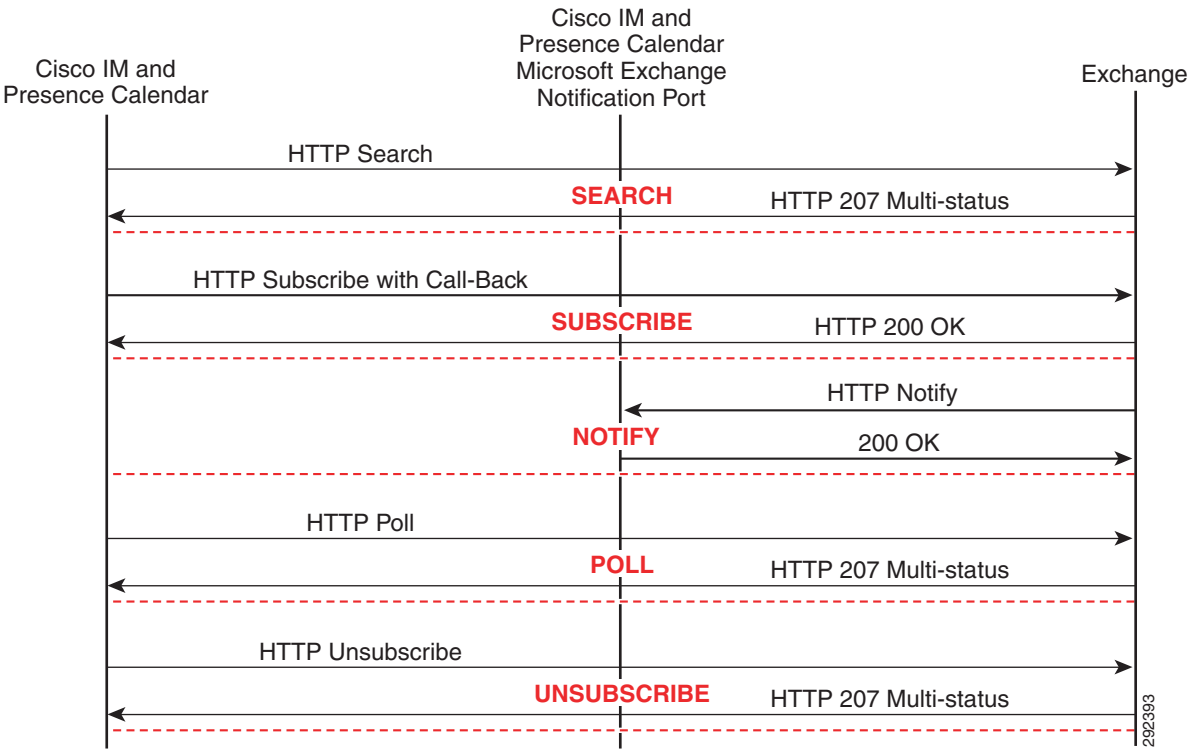
Cisco IM and Presence Service has the ability to retrieve calendar state and aggregate it into a presence status via the calendar module interface with Microsoft Exchange 2010 or 2013 server side integration. Cisco does not provide configuration, deployment, or best practice procedures for Microsoft Exchange, but Cisco does provide the guidelines listed in this section for integrating Cisco IM and Presence Service with the calendar module interface of Microsoft Exchange 2010 or 2013.

Microsoft Exchange integration is supported with Microsoft Active Directory 2008 and Active Directory 2012 as well as Windows Server 2008 and Windows Server 2012. Microsoft Exchange 2010 or 2013 makes the calendar data available from the server through Exchange Web Services (EWS), which allows submitting requests and receiving notifications from Microsoft Exchange. The integration with Microsoft Exchange is done through a separate Presence Gateway configuration for calendar applications. Once the administrator configures a Presence Gateway for Outlook, the user has the ability to enable or disable the aggregation of calendar information into their presence status.

The exchange ID that is used to retrieve calendar information is taken from the email ID of the LDAP structure for that user. If the email ID does not exist or if LDAP is not being used, then the Cisco IM and Presence Service user ID is mapped as the exchange ID.

Information is gathered via a subscription for calendar state from the Cisco IM and Presence Service to the Microsoft Exchange server. Figure 20-27 depicts this communication.

Figure 20-27 Outlook Web Access Communication Between Cisco IM and Presence Service and Microsoft Exchange



Microsoft Outlook Calendar Integration

The IM and Presence Service can incorporate Microsoft Outlook Calendar free and busy data when publishing a user's availability. This feature helps users automatically maintain their availability and status information. Because it is based on a server-to-server integration, it is available to other users whether or not the originating user is logged in. The Microsoft Outlook Calendar feature requires the establishment of a gateway connection to the Microsoft Exchange Server and is compatible with Microsoft Exchange Servers 2003, 2007, and 2010.

Note Cisco IM and Presence Service can be deployed with a single Microsoft Exchange Server or with multiple Microsoft Exchange Servers, in a single forest only. Microsoft Exchange deployment allows for clustering of multiple Exchange servers; therefore, Cisco IM and Presence Service will honor the REDIRECT message to the exchange server that is hosting the user for which Cisco IM and Presence Service is requesting status.

Multi-Language Calendar Support

In cases where the requirements for a calendar integration deployment specify more than one language, use the following design guidelines:

- Cisco IM and Presence Service, as well as Cisco Unified Communications Manager, must have the appropriate locales installed for the users to select their locale.
- Cisco IM and Presence Service supports all the standard Unified Communications locales for calendar integration.
- Users must be configured for the locale that is desired, either through the end user pages or administratively through the Bulk Administration Tool.
- Cisco IM and Presence Service sends the appropriate locale folder with the initial query. Queries are redirected, if required, through the response of the initial Front-End or Client Access Microsoft Exchange server.

Exchange Web Services Calendar Integration

Cisco IM and Presence Service can be configured to allow for Microsoft Exchange Web Services to collect calendar state information to be aggregated into an overall presence view of the user. If the users mailbox is located on the configured Exchange server, Cisco IM and Presence Service will communicate directly with the Exchange server; whereas, if the users mailbox is located on a different Exchange server than the one configured, Cisco IM and Presence Service will follow the Exchange server redirection to find the server where the users mailbox is located. Only Exchange Servers from the server farm can serve as the configured Exchange server, and you are required to specify only one of these servers from the server farm.

Microsoft Exchange Web Services specifies the protocol used to transact with the Exchange Client Access Servers independent of the language that the end-user uses; therefore, there is no need to utilize the locale to determine the language of the end-user. Cisco IM and Presence Service calendar integration is supported with a single Microsoft Exchange forest only.

Cisco IM and Presence Exchange Web Services calendar integration supports both a polling of calendar information as shown in Figure 20-28 as well as a subscription/notification for calendar information as shown in Figure 20-29. Various configuration parameters control the rate of polling intervals, the frequency of subscriptions, and the fault tolerance of timers. For additional configuration details, refer to the *Integration Note for Configuring Cisco IM and Presence with Microsoft Exchange*, available at

https://www.cisco.com/en/US/products/ps6837/products_installation_and_configuration_guides_list.html

Figure 20-28 Exchange Web Services Polling with Cisco IM and Presence Service Calendar

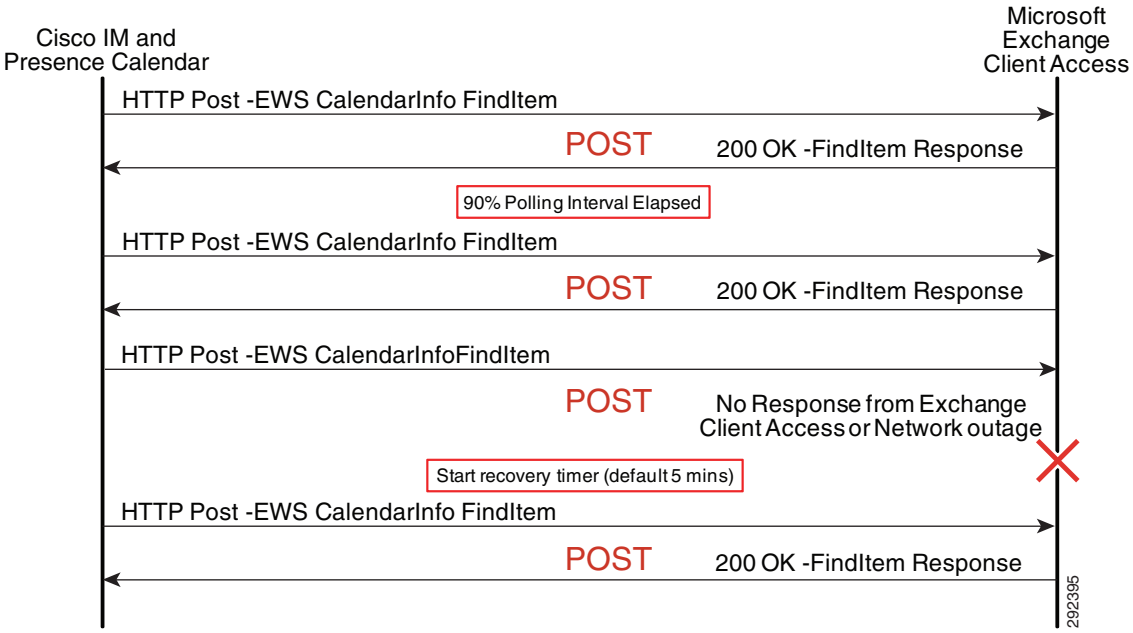
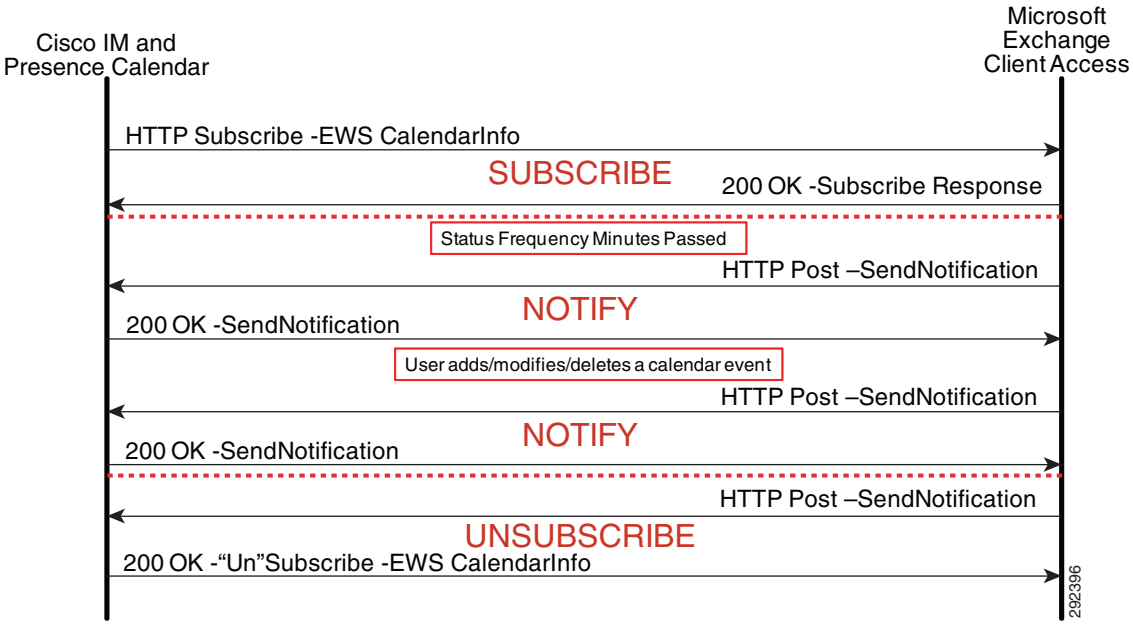


Figure 20-29 Exchange Web Services Subscription/Notification with Cisco IM and Presence Service Calendar



Exchange Web Services Auto Discover is also supported by Cisco IM and Presence Service if a service connection point (SCP) Active Directory object has been created for each server where the Client Access Server (CAS) role is installed. The calendar gateway is configured with Auto Discover using the domain

and optionally the site instead of a host and port. Cisco IM and Presence Service uses the auto-discover algorithm to determine which Exchange Web Services URL to use in contacting the correct Client Access Server Exchange Server.

On-Premises Cisco IM and Presence Service Mobility Integration

Cisco IM and Presence Service has the ability to integrate contact lists and presence state with Cisco Jabber Mobile IM. Jabber Mobile IM continues to communicate directly with Cisco Unified CM, while Cisco Unified CM communicates with Cisco IM and Presence Service via AXL/SOAP and SIP.

An application user must be configured on Cisco IM and Presence Service and Cisco Unified CM before Cisco Unified CM can establish an administrative session with Cisco IM and Presence Service. Cisco Jabber Mobile IM end-user logins will generate a Cisco Unified CM SOAP request to Cisco IM and Presence Service for system configuration, user configuration, contact list, presence rules, and application dial rules, followed by Unified Communicator Change Notifier (UCCN) configuration and Presence SIP subscriptions.

On-Premises Cisco IM and Presence Service Third-Party Open API

Cisco IM and Presence Service has the ability to integrate with third-party applications through HTTP in addition to SIP/SIMPLE and XMPP. The HTTP interface has a configuration interface as well as a presence interface via Representational State Transfer (REST). The Third-Party Open API provides two mechanisms to access presence: a real-time eventing model and a polling model.

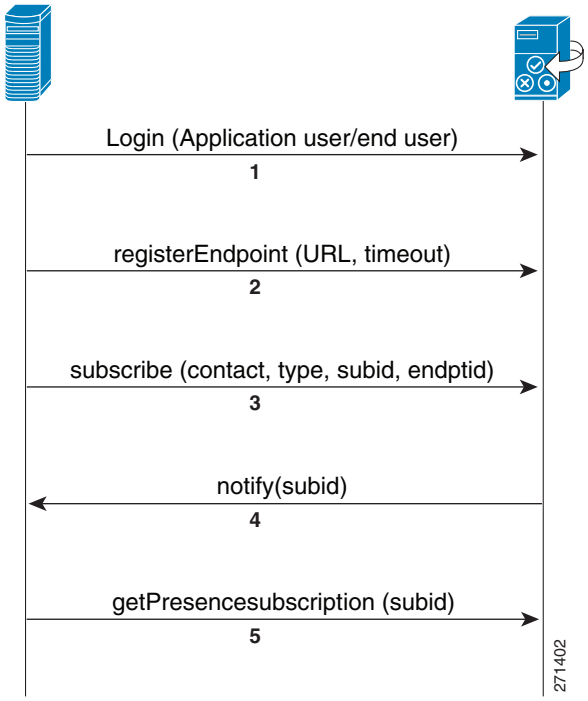
For more information on the Third-Party Open API, refer to the Cisco Developer Community at

https://developer.cisco.com/web/cdc

Real-Time Eventing Model

The real-time eventing model uses an application user on Cisco IM and Presence Service to establish an administrative session, which allows for end users to log in with that session key. Once the end user has logged in, the user registers and subscribes for presence updates using Representational State Transfer (REST). Figure 20-30 highlights the Third-Party Open API real-time eventing model interaction with Cisco IM and Presence Service.

Figure 20-30 Third-Party Open API Real-Time Eventing Model



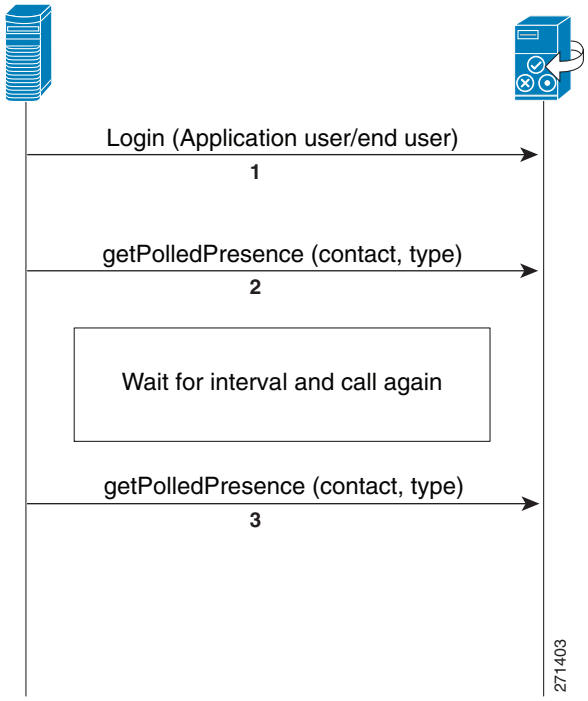
The call flow in Figure 20-30 illustrates the following sequence of events:

1. The application initiates a SOAP login request to Cisco IM and Presence Service via the super-user application user (APIUser), and Cisco IM and Presence Service returns a session key. The application can then log in the end-user with this session key (essentially, the end-user logs in via the application).
2. The end user registers the endpoint using the application-user session key.
3. The application initiates a subscribe request (using the session key) on behalf of the end user to retrieve user information, contact list, and presence rules.
4. Cisco IM and Presence Service sends a notification – unsecured.
5. The application requests the user's presence status.

Polling Model

The polling model uses an application user on Cisco IM and Presence Service to establish an administrative session, which allows for end users to log in with that session key. Once the end user has logged in, the application requests presence updates periodically, also using Representational State Transfer (REST). Figure 20-31 highlights the Third-Party Open API polling model interaction with Cisco IM and Presence Service.

Figure 20-31 Third-Party Open API Polling Model



The call flow in Figure 20-31 illustrates the following sequence of events:

1. The application initiates a SOAP login request to Cisco IM and Presence Service via the super-user application user (APIUser), and Cisco IM and Presence Service returns a session key. The application can then log in the end-user with this session key (essentially, the end-user logs in via the application).
2. The application requests presence state and bypasses the eventing model.
3. The application requests presence state and bypasses the eventing model.

Note Both Basic presence and Rich presence can be retrieved; however, the polling model puts an additional load on the presence server.

Extensible Messaging and Presence Protocol Interfaces

The XCP architecture allows for two additional open interfaces for presence, instant messaging, and roster management: a client XMPP interface and a Cisco AJAX XMPP Library interface. The client XMPP functionality enables third-party XMPP clients to integrate presence, instant messaging, and roster management, and it is a complementary interface to the SIP/SIMPLE interface on Cisco IM and Presence Service. The client XMPP interface is treated as a normal XMPP client within Cisco IM and Presence Service; therefore, sizing of the interface should be treated as a normal XMPP client.

The Cisco AJAX XMPP Library API provides a Web 2.0 style of interface to integrate XCP features into web applications and widgets, and it is made directly available from Cisco IM and Presence Service. The Cisco AJAX XMPP Library API is exclusively a client-side JavaScript library that communicates to the Bidirectional-streams Over Synchronous HTTP (BOSH) interface, which is essentially an XMPP over HTTP interface that allows the server to push data to a web browser through a long-polling technique.

Observe the following requirements when integrating either model of the Third-Party Open API with Cisco IM and Presence Service:

- Certificates are required for the presence interface (sipproxy.der) and the configuration interface (tomcat_cert.der).
- No more than 1000 Third-Party Open API users can be integrated per Cisco IM and Presence Service deployment.
- To improve performance, balance the Third-Party Open API users across all servers in the Cisco IM and Presence Service cluster.

You can obtain additional information and support for use of the Cisco IM and Presence Service Third-Party Open API through Cisco Developer Services, available at:

https://developer.cisco.com/web/cupapi

Information and assistance for developers is also available from the Cisco Developer Community, which is accessible through valid Cisco login authentication at:

https://developer.cisco.com/

Design Considerations for On-Premises Cisco IM and Presence Service

- If LDAP integration is possible, LDAP synchronization with Unified CM should be used to pull all user information (number, ID, and so forth) from a single source. However, if the deployment includes both an LDAP server and Unified CM that does not have LDAP synchronization enabled, then the administrator should ensure consistent configuration across Unified CM and LDAP when configuring user directory number associations.
- Cisco IM and Presence Service marks Layer 3 IP packets via Differentiated Services Code Point (DSCP). Cisco IM and Presence Service marks all IM and Presence traffic based on the Differential Service Value service parameter under SIP Proxy, which defaults to a value of DSCP 24 (PHB CS3).
- Presence Policy for Cisco IM and Presence Service is controlled strictly by a defined set of rules created by the user.
- Use the service parameter IMP PUBLISH Trunk to streamline SIP communication traffic with the Cisco IM and Presence Service.
- Associate presence users in Unified CM with a line appearance, rather than just a primary extension, to allow for increased granularity of device and user presence status. When using the service parameter IMP PUBLISH Trunk, you must associate presence users in Unified CM with a line appearance.
- A Presence User Profile (the user activity and contact list contacts and size) must be taken into consideration for determining the server hardware and cluster topology characteristics. The Cisco IM and Presence system architecture is based on an average contact list size of 75 contacts per user on a fully populated system. While per-user contact list size will vary across the system, if significant numbers of users on the system exceed the average list size of 75 contacts, system performance will be impacted. By default the maximum contact list size is 200. If some users will exceed 200 contacts, this maximum contact list size can be changed by modifying the Presence Settings of the IM and Presence cluster. For additional sizing information, see the sections on Cisco IM and Presence, page 25-33, and Roster Management, page 25-34.
- Use the User Assignment Mode for Presence Server enterprise parameter default of **balanced** for best overall cluster performance.

- Cisco IM and Presence Service requires an external database instance for each server in the cluster for persistent chat, and one database instance per cluster for message archiving. Third-party compliance supports mapping of all or a subset of servers in an IM and Presence Service cluster to one external compliance database. The three external database options are flexible deployments wherein there can be multiple compliance servers per IM and Presence server, multiple IM and Presence servers per compliance server, or a combination thereof.
- Compliance servers can be dedicated or shared with IM and Presence nodes in the same cluster. Cisco recommends deploying two compliance servers per sub-cluster pair, with both IM and Presence nodes defined in the compliance profile. One compliance server for the entire cluster is also supported but not recommended.
- You can install the database on either a Linux or a Windows operating system. Refer to the relevant database documentation for details on the supported operating systems and platform requirements:
  - PostgreSQL documentation available at https://www.postgresql.org/docs/manuals/
  - Oracle documentation available at https://docs.oracle.com/en/database/database.html
- Cisco IM and Presence Service supports a total of 75,000 users per cluster for full Unified Communications mode. The sizing for users must take into account the number of SIP/SIMPLE users and the number of XMPP users. XMPP users have slightly better performance because SIP/SIMPLE users employ the IM Gateway functionality into the XCP architecture.
- All eXtensible Communications Platform (XCP) communications and logging are stored in GMT and not localized to the installed location.
- For ease of user migration and contact list migration, Cisco IM and Presence Bulk Administration Tool supports bulk contact list importation using a comma-separated value (csv) file as input for this bulk importation.

For a complete listing of ports used by Cisco IM and Presence Service, refer to *Port Usage Information for Cisco IM and Presence*, available at

https://www.cisco.com/en/US/products/ps6837/products_device_support_tables_list.html

Contact and Watcher List Recommendations

Note The guidelines provided in this section are based strictly on what has been validated by Cisco under specific test conditions. The recommendations are intended to help guide the deployment and distribution of presence users in an IM and Presence deployment with regard to managing contact lists and watcher lists in the cluster.

IM and Presence Cluster

The IM and Presence standard deployment is designed to support 45,000 presence-enabled users in a fully loaded cluster across 3 IM and Presence sub-cluster pairs (6 nodes) and deployed using15k-User IM and Presence VM templates.

The recommendation for contact lists and watcher lists is not to exceed a cluster average of 75 presence-enabled contacts per user in the cluster. This value is derived from validation tests of 100 total contacts per user on average, split with 75 presence-enabled users and 25 non-presence users in their respective buddy lists. The rest of this discussion focuses on the 75 presence-enabled users because that is the factor that impacts system performance and scalability the most.

To help manage the IM and Presence system performance and stability, it is crucial to properly maintain, manage, and monitor the IM and Presence roster tables, which basically consist of end-user contact lists and watcher lists for all presence users.

Note The cluster average of 75 presence-enabled contacts per user is not the same as the service parameter(s) for **Maximum Contact List Size (per user)** and **Maximum Watchers (per user)**, where the default values are 150 and 200 respectively, depending on Cisco IM and Presence release version.

Presence State Change Impact

To help understand how a state change impacts the system, consider the events from the perspective of a single user on a typical work day. Assume that user, named Bob, has 75 contacts in his buddy list and those 75 contacts also have Bob as a contact in their buddy lists (although that is not necessarily the case in most deployments).

When Bob first logs in to his desktop Jabber client, 75 notifications are generated and sent out to all the contacts in his buddy list. If Bob then uses his desk phone to make a call, another 75 notifications are sent to his contacts to update his status to "on the phone." Every status change for Bob initiates a notification per contact in his buddy list *if* Bob is also in that contact's buddy list.

All presence-enabled users initiate status change updates to all presence-enabled contacts in their buddy lists. Status changes include actions such as picking up the phone handset, making a call, hanging up, joining a meeting, and so on. So for every user with 75 presence-enabled contacts, every action generates 75 notifications at minimum.

For example, consider a deployment of 3,000 presence-enabled users, each of which has 75 contacts in their buddy list. If all 3,000 users join a conference or all-hands meeting at the same time, that would generate 225,000 presence update notifications. Similarly, in a deployment of 9,000 presence users, if all users perform actions that cause presence status changes at the same time, that would generate 675,000 presence update notifications. These notifications utilize system resources and impact system performance, especially during peak usage times. That is why it is crucial to distribute users evenly across the cluster and to avoid exceeding the recommended average of 75 contacts per user.

Distribution of Presence Users

The budget, or maximum number of contact entries allowed, for the IM and Presence database can be calculated from the total number of configured presence users multiplied by the recommended average number of presence contacts (75) that each user can have in their buddy list. For example, the recommended budget (maximum number of IM and Presence contacts) for a fully loaded cluster of 45,000 users would be:

(45,000 users) * (75 contacts per user) = 3.375M IM and Presence contacts for the cluster

If there are 3 sub-cluster pairs (6 nodes) in the deployment with 15,000 users in each sub-cluster, then on average there would be:

15,000 users * 75 contacts per user = 1.125M contacts per sub-cluster pair

If each IM and Presence user has no more than 75 contacts in their buddy list, then the users can be distributed equally across the sub-clusters, as indicated in the above calculation. However, if some users require more than 75 contacts in their buddy lists, then a custom distribution is required, as described in the following section.

Tip Make sure that the users with a high number of contacts (more than 75) do not all reside on the same IM and Presence node and, instead, are distributed evenly across the nodes.

Custom Distribution

Again consider a fully loaded IM and Presence cluster with 45,000 users distributed across three sub-clusters. Assume, for example, that 1,000 users in the cluster need to have 500 presence-enabled contacts in their buddy lists. In this case those 1,000 users would require:

(1,000 users) * (500 contacts per user) = 500,000 IM and Presence database entries

The maximum number of IM and Presence database entries allowed for the entire cluster is 3.375M. After accounting for the 1,000 users with 500 contacts each, the remaining available contact entries would be:

3,375,000 – 500,000 = 2,875,000 contact entries available for the remaining 44,000 presence users in the cluster.

If the available contact entries are distribute evenly among the remaining 44,000 users in the cluster, that would provide:

2,875,000/44,000 = maximum of 65 contacts per user

Note In the above example, the users with 500 contacts must be distributed evenly across the three sub-cluster pairs. It is also critical to ensure system stability by frequently monitoring and adjusting the distribution of IM and Presence users as their usage and resource requirements change. In addition, if the number of contacts required by any user in the cluster is higher than the default value of the service parameter **Maximum Contact List Size (per user)**, then that parameter setting and the setting for **Maximum Watchers (per user)** must both be changed to the higher value.

Status Change Notify Impact

From the example above, the custom distribution with 500 contacts per user would generate 500 notifications for every status change compared to 75 notifications per status change with the evenly distributed option. The following examples illustrate the impact comparison between 1000 users with 75 contacts and 500 contacts:

- 1000 * 75 = (75,000 notifications) / (3,600 seconds, or 1 hour) = 21 notifications/sec per cluster, or 7 notifications/sec per sub-cluster pair.
- 1000 * 500 = (500,000 notifications) / (3,600 seconds, or 1 hour) = 139 notifications/sec per cluster, or 47 notifications/sec per sub-cluster pair.

Mobile and Remote Access

Users can access their collaboration tools from outside the corporate firewall without a VPN client. Using Cisco collaboration gateways, the client can connect securely to your corporate network from remote locations such as public Wi-Fi networks or mobile data networks.

Cisco Unified Communications mobile and remote access is a core part of the Cisco Collaboration Edge Architecture. It allows endpoints such as Cisco Jabber to have their registration, call control, provisioning, messaging, and presence services provided by Cisco Unified Communications Manager (Unified CM) when the endpoint is not within the enterprise network. Cisco Expressway provides secure firewall traversal and line- side support for Unified CM registrations.

Note that third-party SIP or H.323 devices can register to a Cisco VCS connected via a neighbor zone to a Cisco Expressway and, if necessary, interoperate with Unified CM registered devices over a SIP trunk.

For information on how to set up the Cisco Expressway servers, refer to the latest version of the *Cisco Expressway Basic Configuration Deployment Guide* and the *Mobile and Remote Access via Cisco Expressway Deployment Guide*, both available at

https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html

When using mobile and remote access, the following Jabber features are not supported:

- Directory access mechanisms other than UDS
- Certificate provisioning to remote endpoints
- File transfer
- Deskphone control mode.

Third-Party Presence Server Integration

Cisco IM and Presence Service provides an interface based on SIP and SIP for Instant Messaging and Presence Leveraging Extensions (SIMPLE) for integrating SIP and SIMPLE applications into the Cisco Unified Communications solution. You can configure and integrate a third-party presence server or application with this SIP/SIMPLE interface to provide presence aggregation and federation.

Microsoft Communications Server for Remote Call Control (RCC)

For all setup, configuration, and deployment of Microsoft products, refer to the documentation at:

https://www.microsoft.com/

Cisco does not provide configuration, deployment, or best practice procedures for Microsoft Communications products, but Cisco does provide the guidelines listed below for integrating Cisco IM and Presence Service with Microsoft Lync.

Cisco Systems has developed documentation to describe feature interoperability and configuration steps for integrating Cisco IM and Presence Service with Microsoft Lync. You can access this documentation at:

https://www.cisco.com/en/US/products/ps6837/products_installation_and_configuration_guides_list.html

Guidelines for Integrating Cisco IM and Presence Service with Microsoft Lync

The following guidelines apply when integrating the Cisco IM and Presence Service and Microsoft Lync:

- Communications between Cisco IM and Presence Service and Microsoft Lync uses the SIP/SIMPLE interface. However, Microsoft Lync tunnels Computer-Supported Telecommunications Applications (CSTA) traffic over SIP. Therefore, the CTI gateway on the Cisco IM and Presence Service must be configured to handle the CSTA-to-CTI conversion for Click to Call phone control.
- Cisco IM and Presence Service deployment with Microsoft Lync for Remote Call Control, should consist of a single subcluster pair of servers that make up the Cisco IM and Presence Service cluster.
- The following table lists the number of users supported per platform. The user count is based solely on the Unified CM platform equivalent, regardless of the IM and Presence Service platform.

Cisco Unified Communications Manager VM Configuration Template	Number of Microsoft Office Communicator or Lync Users Supported per Server	Number of Microsoft Office Communicator or Lync Users Supported per Cluster
1,000 User	1,000	4,000
2,500 User	2,500	10,000
7,500 User	7,500	30,000
10,000 User	10,000	40,000

- You must configure the same end-user ID in LDAP, Unified CM, and Microsoft Lync. This practice avoids any conflicts between Microsoft Lync authentication with Active Directory (AD) and the end-user configuration on Unified CM, as well as conflicts with user phone control on Unified CM.

  For Active Directory, Cisco recommends that the user properties of General, Account, and Communications all have the same ID. To ensure the Cisco IM and Presence Service users are consistent, LDAP Synchronization and Authentication should be enabled with Unified CM.
- You must configure Microsoft Lync Host Authentication to contain the Cisco IM and Presence Service publisher and subscriber.
- You can configure routing of the SIP messages to Cisco IM and Presence Service by means of Static Routes in the Microsoft Lync properties.
- You must configure an incoming and outgoing access control list (ACL) on the Cisco IM and Presence Service to allow for communications with Microsoft Lync.
- You must enable each user for use of Microsoft Lync in the Cisco IM and Presence Service configuration, in addition to enabling each user for presence in Unified CM.
- Take into account bandwidth considerations for Microsoft Lync login due to the exchange of configuration information between Microsoft Lync and the Microsoft Communications Server, and due to initial communication with the Cisco IM and Presence Service CTI gateway.
- To address the issue of a reverse look-up of a directory number that corresponds to a user, use the guidelines documented in the *Release Notes for Cisco IM and Presence*, available at

  https://www.cisco.com/en/US/products/ps6837/prod_release_notes_list.html

In-the-Cloud Service and Architecture

This section describes the in-the-cloud service and architecture for Cisco IM and Presence Service. This hosted service provides the same user experience as the on-premises solution.

Cisco WebEx Messenger

Cisco WebEx Messenger is a multi-tenant software-as-a-service (SaaS) platform for synchronous and asynchronous collaboration. The WebEx Messenger platform is hosted inside the Cisco WebEx Collaboration Cloud and it enables collaborative applications and integrations, which allows for organizations and end users to customize their work environments. For additional information on the Cisco WebEx Messenger service, refer to the documentation available at

https://developer.cisco.com/web/webex-developer

For more information on the Cisco Collaboration Cloud, refer to the documentation available at

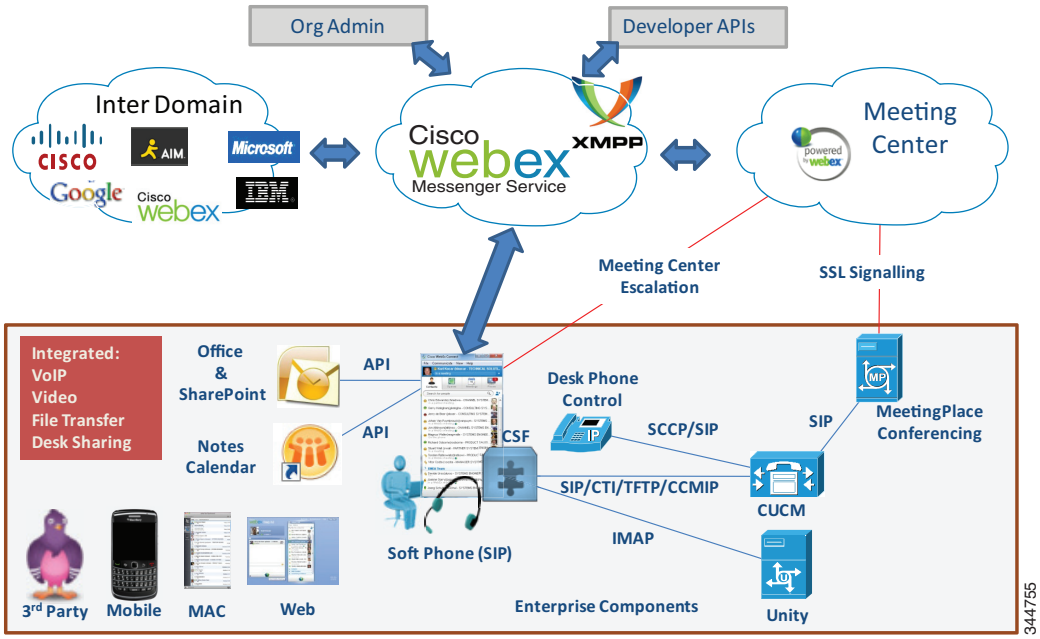
https://www.cisco.com/en/US/solutions/ns1007/collaboration_cloud.html

Deploying Cisco WebEx Messenger Service

A Cisco WebEx Messenger solution deployment consists of the following components, as depicted in Figure 20-32:

- A secure connection (SSL and AES) to the Cisco WebEx Messenger XMPP cloud platform for presence, instant messaging, VoIP, PC-to-PC video, media transfer (screen capture and file transfer), and desktop sharing
- Cisco WebEx Meetings
- XMPP federation with other WebEx Messenger organizations and third-party XMPP clients and XMPP instant messaging (IM) networks
- Cisco Unified Communications integration for call control, voice messaging, and call history
- Microsoft Outlook and IBM Lotus Notes calendar integration
- Integration to Microsoft Outlook for presence and click-to-communicate functionality

Figure 20-32 Deploying Cisco WebEx Messenger Service



Centralized Management

Cisco WebEx Messenger service provides a web-based administrative tool to manage the solution across the organization. Cisco WebEx Messenger service users are configured and managed through the Cisco WebEx Administration Tool, which enables administrators to set up basic security and policy controls for features and services. These policies can be applied enterprise-wide, by group, or individually. There are various methods to provision the user database that are further described in the Cisco WebEx administrator's guide available at

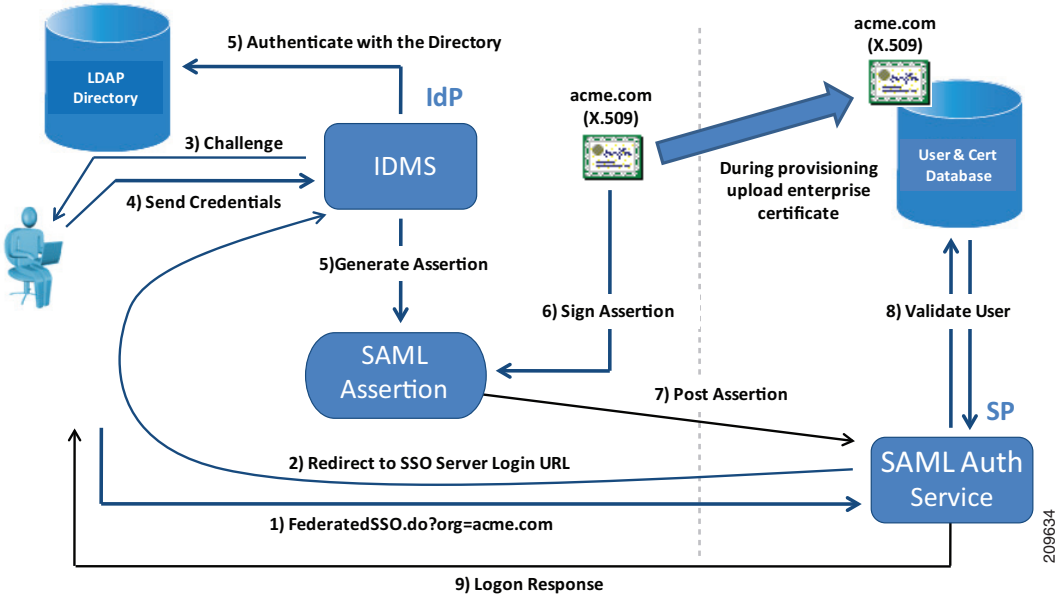
https://www.webex.com/webexconnect/orgadmin/help/index.htm

Single Sign On

Single Sign On (SSO) enables companies to use their on-premises SSO system, including Security Assertion Markup Language (SAML) support, to simplify the management of Cisco WebEx Messenger or IM and Presence Service by allowing users to securely log into any of the Unified Communications applications in the solution using their corporate login credentials. The user's login credentials are not sent to Cisco, thus protecting the user's corporate login information. Figure 20-33 shows the credential handshake that occurs on user login to Cisco WebEx Messenger as well as Unified CM.

Note If Cisco Jabber is deployed with Cisco WebEx Meeting Server, Cisco Unified CM and WebEx Meeting Server must be in the same domain.

Figure 20-33 User Login Authentication Process for Cisco WebEx Messenger Service



A user account can be configured to be created automatically the first time a user logs into Cisco IM client. Users are prevented from accessing the Cisco WebEx Messenger service if their corporate login account is deactivated.

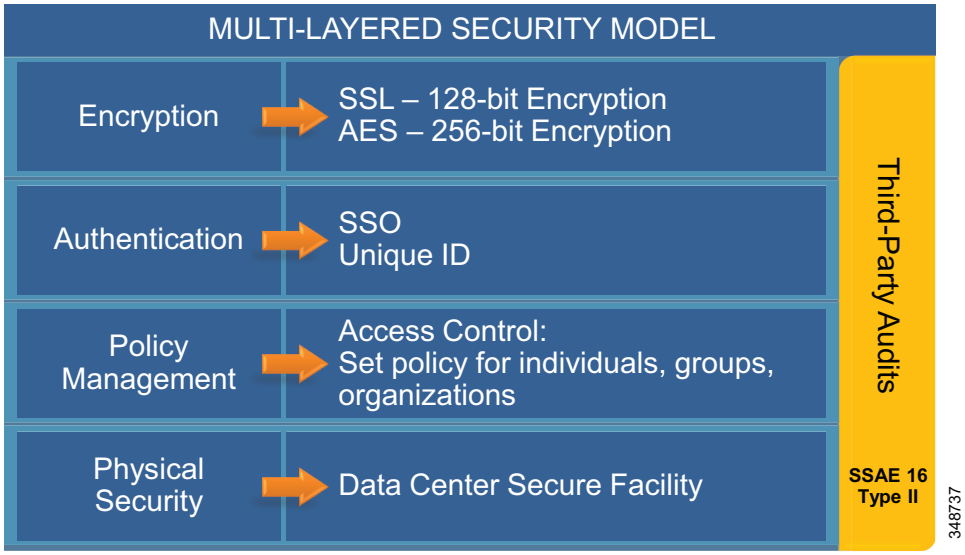
For more information on Single Sign On with WebEx Messenger service, refer to the documentation available at

https://developer.cisco.com/web/webex-developer/sso-reference

Security

The Cisco WebEx security model consists of functional layers of security. Figure 20-34 illustrates the separate but interrelated elements that compose each layer.

Figure 20-34 WebEx Security Model



The bottom layer represents the physical security in the Cisco WebEx data centers. All employees go through an extensive background check and must provide dual-factor authentication to enter the datacenter.

The next level is policy management, where the WebEx Messenger organization administrator can set and manage access control levels by setting different policies for individual users, groups, or the entire Cisco WebEx Messenger organization. White-list policies, specific to external users or domains, can be created to allow instant messaging exchanges. The Cisco WebEx Messenger organizational model also allows for the creation of specific roles and groups across the entire user base, which allows the administrator to assign certain privileges to roles or groups as well as to set policies, including access control, for the entire organization.

Access to the Cisco WebEx Messenger service is controlled at the authentication layer. Every user has a unique login and password. Passwords are never stored or sent over email in clear text. Passwords can be changed only by the end-users themselves. The administrator can choose to reset a password, forcing the end-user to change his or her password upon the next login. Alternatively, an administrator may choose to use the Single Sign On (SSO) integration between Cisco WebEx Messenger service and the company's directory to simplify end-user access management. The Single Sign On integration is achieved through the use of an Identity Management System (IDMS).

The encryption layer ensures that all instant messaging communications between Cisco WebEx Messenger users is encrypted. All instant messaging communication between Cisco WebEx Messenger users and the server in the Messenger Collaboration cloud is encrypted by default using SSL encryption. An additional level of security is available whereby IM communication can be encrypted end-to-end using 256-bit AES level encryption.

The Cisco WebEx Messenger platform uses third-party audits such as the SSAE 16 Type II audit to provide customers with an independent semi-annual security report. This report can be reviewed by any customer upon request with the Cisco Security organization. For additional Cisco WebEx Messenger service security, refer to the C*isco WebEx Connect Security White Paper*, available at

https://www.cisco.com/en/US/products/ps10528/prod_white_papers_list.html

Firewall Domain White List

Access control lists should be set specifically to allow all communications from the webex.com and webexconnect.com domains and all sub-domains for both webex.com and webexconnect.com. The WebEx Messenger platform sends email to end-users for username and password communications. These email messages come from the mda.webex.com domain.

Logging Instant Messages

Cisco WebEx Messenger service instant messaging communications are logged on the local hard drive of the personal computer where the user is logged in. Instant message logging is a capability in Cisco WebEx Messenger service that can be enabled by means of policy through the Org Admin tool.

The end-user can set logging specifics, whether to enable or disable logging, and how long the logs are kept. These message history settings are located under General in the IM client preferences.

Customers looking for advanced auditing and e-discovery capabilities should consider third-party solutions. Currently Cisco does not provide support for advanced auditing of instant messaging communications. Cisco WebEx Messenger service, however, does allow for logging and archiving of instant messages exchanged between users. Archiving of the logs is possible though the use of third-party SaaS archiving services, or the logs can be delivered securely to an on-premises SMTP server.

For additional information on instant message archiving, refer to the Cisco WebEx administrator's guide available at

https://www.webex.com/webexconnect/orgadmin/help/index.htm

Capacity Planning for Cisco WebEx Messenger Service

A single end-user requires only a 56 kbps dial-up Internet connection to be able to log in to WebEx Messenger service and get the basic capabilities such as presence, instant messaging, and VoIP calling. However, for a small office or branch office, a broadband connection with a minimum of 512 kbps is required in order to use the advanced features such as file transfer, screen capture, and PC-to-PC video calling. For higher quality video such as High Definition 720p, the minimum bandwidth connection recommendation is 2 Mbps.

For additional information on network and desktop requirements, refer to the Cisco WebEx administrator's guide available at

https://www.webex.com/webexconnect/orgadmin/help/index.htm

Cisco Webex Messenger deployment network requirements are available at

https://www.webex.com/webexconnect/orgadmin/help/17161.htm

High Availability for Cisco WebEx Messenger Service

WebEx Messenger is a Software-as-a-Service (SaaS) application. The end-user device must be connected to the Internet for the end user to log in to the IM client. A standard Internet connection is all that is required. If an end user is remote, it is not necessary for that user to be connected through the company VPN in order to log in to the WebEx Messenger service. Cisco WebEx Messenger service IM clients can be deployed in a highly available redundant topology. Deployment of the Cisco WebEx Messenger Software-as-a-Service architecture consists of various network and desktop requirements described in this section.

High Availability

With the use of the multi-tenant Software-as-a-Service architecture, if any individual server in a group fails for any reason, requests can be rerouted to another available server in the Cisco WebEx Messenger Platform.

The Cisco WebEx Network Operations Team provides 24x7 active monitoring of the Cisco WebEx Collaboration Cloud from the Cisco WebEx Network Operations Center (NOC). For a comprehensive overview of the Cisco WebEx technology, refer to the information at

https://www.cisco.com/en/US/solutions/ns1007/collaboration_cloud.html

Redundancy, Failover, and Disaster Recovery

The Cisco WebEx Global Site Backup architecture handles power outages, natural disaster outages, service capacity overload, network capacity overload, and other types of service interruptions. Global Site Backup supports both manual and automatic failover. The manual failover mode is typically used during maintenance windows. The automatic failover mode is used in case of real-time failover due to a service interruption.

Global Site Backup is automatic and transparent to the end users, it is available for all users, and it imposes no limits on the number of users that can fail-over.

Global Site Backup consists of the following main components:

- Global Site Service — Is responsible for monitoring and switching traffic at the network level.
- Database Replication — Ensures that the data transactions occurring on the primary site are transferred to the backup site.
- File Replication — Ensures that any file changes are maintained in synchronization between the primary and the backup site.

Design Considerations for Cisco WebEx Messenger Service

Cisco WebEx Messenger is deployed as a Software-as-a-Service model, therefore design and deployment considerations are minimal. The Cisco WebEx Messenger solution has client options available for the Windows and Mac desktop as well as the popular mobile devices.

Third-Party XMPP Clients Connecting to Cisco WebEx Messenger Service

Although Cisco does not officially support any other XMPP clients to connect to the Cisco WebEx Messenger Service, the nature of the XMPP protocol is to allow end users to connect to presence clouds with various XMPP clients using their WebEx Messenger service credentials. A list of XMPP software clients is available at

https://xmpp.org/software/clients.shtml

Organization policies cannot be enforced on third-party XMPP clients, and features such as end-to-end encryption, desktop share, video calls, PC-to-PC calls, and teleconferences are not supported with third-party clients. To allow non-WebEx Messenger service XMPP IM clients to authenticate to your WebEx Messenger service domain(s), DNS SRV records must be updated. The specific DNS SRV entry can be found in Cisco WebEx administration, under Configuration and IM Federation.

The use of non-Messenger service XMPP clients in Cisco WebEx administration, under Configuration and XMPP IM Clients, must be explicitly allowed.

For additional information on enabling third-party XMPP clients to connect to the WebEx Messenger platform, refer to the Cisco WebEx administrator's guide available at

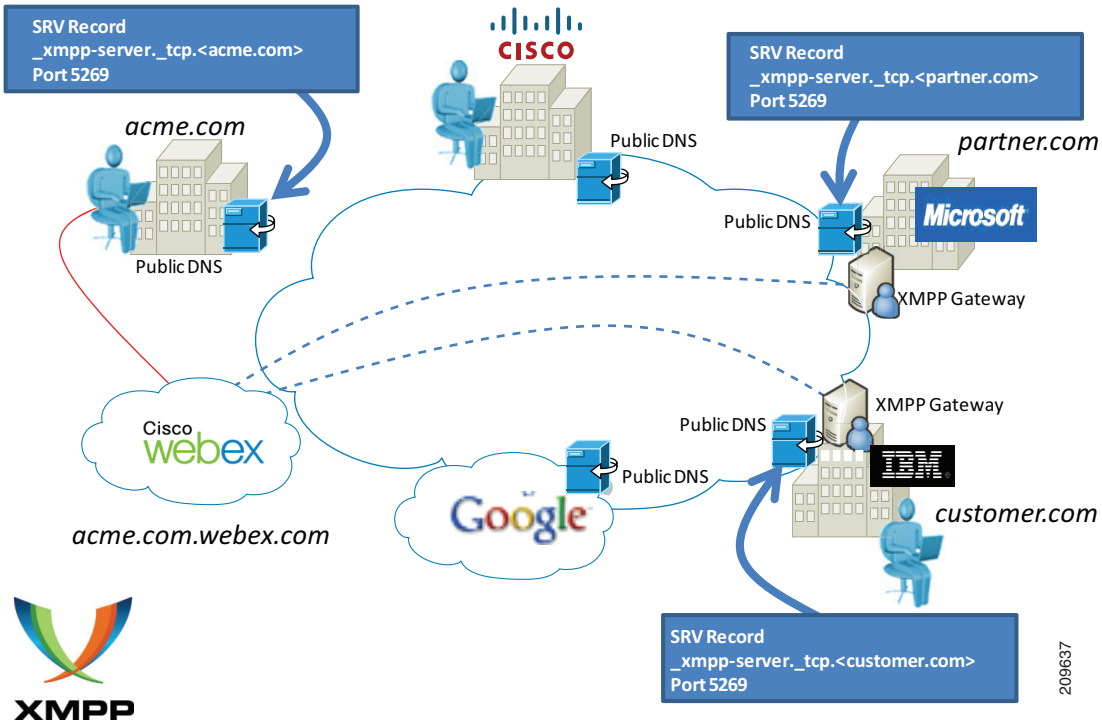
https://www.webex.com/webexconnect/orgadmin/help/index.htm

Instant Messaging and Presence Federation Using Third-Party XMPP Clients

The Cisco WebEx Messenger service network can federate with XMPP-based instant messaging networks such as GoogleTalk and Jabber.org. (See Figure 20-35.) A list of public instant messaging networks based on XMPP is available at

https://xmpp.org/

Figure 20-35 Inter-Domain Federation



Currently the WebEx Messenger service does not interoperate with Yahoo! Messenger and Windows Live Messenger, but it can federate with AIM through a federation gateway.

Other Resources and Documentation

The Cisco WebEx administrator's guide is available at

https://www.webex.com/webexconnect/orgadmin/help/index.htm

Mobile Collaboration

Revised: March 1, 2018

Mobile collaboration solutions and applications provide the ability to deliver features and functionality of the enterprise IP communications environment to mobile workers wherever they might be. With mobile collaboration solutions, mobile users can handle business calls on a multitude of devices and access enterprise applications whether moving around the office building, between office buildings, or between geographic locations outside the enterprise. Mobile collaboration solutions provide mobile workers with persistent reachability and improved productivity as they move between, and work at, a variety of locations.

Mobile collaboration solutions can be divide into two main categories:

- Mobility within the enterprise

  This type of mobility is limited to movement of users within enterprise locations.

- Mobility beyond the enterprise

  This type of mobility refers to mobility beyond the enterprise infrastructure and typically involves some form of Internet, mobile voice network, and/or mobile data network traversal.

Mobility within the enterprise is limited to utilization within the network boundaries of the enterprise, whether those boundaries span only a single physical building, multiple physical buildings in close proximity or separated by long distances, or even home offices where network infrastructure is still controlled and managed by the enterprise when it is extended to the home office.

On the other hand, mobility beyond the enterprise involves a bridging of the enterprise infrastructure to the Internet or mobile provider infrastructures and finds users leveraging public and private networks for connectivity to enterprise services. In some cases the lines between these two types of mobility are somewhat blurred, especially in scenarios where mobile devices are connecting back to the enterprise for collaboration services over the Internet or mobile data and mobile voice networks.

Mobility within the enterprise can be divided into three main areas based on feature sets and solutions:

- Campus or single-site mobility

  With this type of enterprise mobility, users move around within a single physical location typically bounded by a single IP address space and PSTN egress/ingress boundary. This type of mobility involves operations and features such as phone movement from one physical network port to another, wireless LAN device roaming between wireless infrastructure access points, and even Cisco Extension Mobility (EM), where users temporarily apply their device profile including their enterprise number to a particular phone in a different area.

- Multisite mobility

  With this type of mobility, users move within the enterprise from one physical location to another, and this movement typically involves crossing IP address spaces as well as PSTN egress/ingress boundaries. This type of mobility involves the same types of operations and features as with campus mobility (physical hardware moves, WLAN roaming, and Cisco Extension Mobility) but replicated at each site within the enterprise. In addition, the Device Mobility feature can be leveraged to ensure that, as user's move devices between sites, phone calls are routed through the local site egress gateway, media codecs are negotiated appropriately, and call admission control mechanisms are aware of the device's location.

- Remote site mobility

  With this type of mobility, users move to a location outside the enterprise but still have some form of secure connection back to the enterprise, which virtually extends the enterprise network to the remote location. This type of mobility involves either VPN-based remote enterprise connectivity or VPN-less remote enterprise connectivity. VPN remote enterprise connectivity includes remote teleworker solutions such as Cisco Virtual Office as well as other remote connectivity methods such as VPN-capable phones and clients and the Office Extend Access Point feature. VPN-less remote enterprise connectivity enables reverse proxy firewall session-based connections, allowing remote endpoints and clients to connect to the enterprise without requiring a VPN tunnel. VPN-less remote connectivity is supported with the Cisco Expressway mobile and remote access feature.

- Cloud and hybrid services mobility

  This type of mobility includes cloud collaboration services and integrations of cloud and on-premises collaboration services. Because this involves delivery of services from the cloud, any device capable of connecting to the Internet can be used to leverage these services. Regardless of whether a user is inside or outside the enterprise, connected to the enterprise or another network, in motion or at rest, they can consume these cloud services.

Mobility beyond the enterprise can be divided into two high-level Cisco solution sets:

- Cisco Unified Mobility

  As part of Cisco Unified Communications Manager (Unified CM), the Cisco Unified Mobility feature suite offers the ability to associate a mobile user's enterprise number to their mobile or remote devices and provides connectivity between the user's fixed enterprise desk phone on the enterprise network and the user’s mobile device on the mobile voice provider network. This type of functionality is sometimes referred to as fixed mobile convergence.

- Cisco Mobile Client Solutions

  Cisco mobile client applications run on dual-mode smartphones and other mobile devices, and they provide access to enterprise collaboration applications and services. Dual-mode phones provide dual radio antennas for connecting to both 802.11 wireless LAN networks and cellular voice and data networks. With a Cisco mobile client deployed on mobile devices, they can be registered to Cisco Unified CM through the enterprise wireless LAN or over the Internet through public or private Wi-Fi hot spots or the mobile data network, and they can in turn leverage the IP telephony infrastructure of the enterprise for making and receiving voice and video calls over IP. In the case of dual-mode

phones, when mobile users are not associated to the enterprise WLAN or securely attached to the enterprise network with these devices, phone calls are made using the mobile voice provider network. In addition to enabling voice and video services for the mobile device, Cisco mobile clients also provide access to other collaboration services such as voice and instant messaging, presence, and enterprise directory access.

The various applications and features discussed in this chapter apply to all Cisco Unified Communications deployment models unless otherwise noted.

This chapter begins with a discussion of mobility features and solutions available within the enterprise infrastructure. It includes an examination of functionality and design considerations for campus or single-site deployments, multisite deployments, and even remote site deployments. This comprehensive set of solutions provides many benefits for mobile workers within the enterprise, including enterprise-class communications and improved productivity regardless of physical location. This discussion of mobility within the enterprise paves the way for examination of mobility solutions beyond the enterprise that leverage the mobile provider and Internet provider infrastructure and capabilities. These solutions enable a bridging of the enterprise network infrastructure and mobile functionality to the provider network infrastructure in order to leverage advanced mobile features and communication flows that can be built on the solid enterprise mobility infrastructure.

This chapter provides a comprehensive examination of mobility architectures, functionality, and design and deployment implications for enterprise collaboration mobility solutions. The analysis and discussions contained within this chapter are organized at a high level as follows:

- Mobility within the Enterprise
  - Campus Enterprise Mobility, page 21-4
  - Multisite Enterprise Mobility, page 21-11
  - Remote Enterprise Mobility, page 21-26
  - Cloud and Hybrid Services Mobility, page 21-34
- Mobility beyond the Enterprise
  - Cisco Unified Mobility, page 21-47
  - Cisco Mobile Clients and Devices, page 21-76

What's New in This Chapter

Table 21-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 21-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Apple Push Notification service (APNs) for Cisco Jabber	Apple Push Notification Service (APNs) for Cisco Jabber for iPhone and iPad, page 21-99	March 1, 2018
OAuth 2.0 with Refresh Token for Cisco Jabber	Cisco Jabber and OAuth with Refresh Login Flow, page 21-101	March 1, 2018

Mobility Within the Enterprise

This section examines mobility features and solutions available within the enterprise. This examination includes discussions related to architecture, functionality, and design and deployment implications for the following types of enterprise mobility

- Campus Enterprise Mobility, page 21-4
- Multisite Enterprise Mobility, page 21-11
- Remote Enterprise Mobility, page 21-26
- Cloud and Hybrid Services Mobility, page 21-34

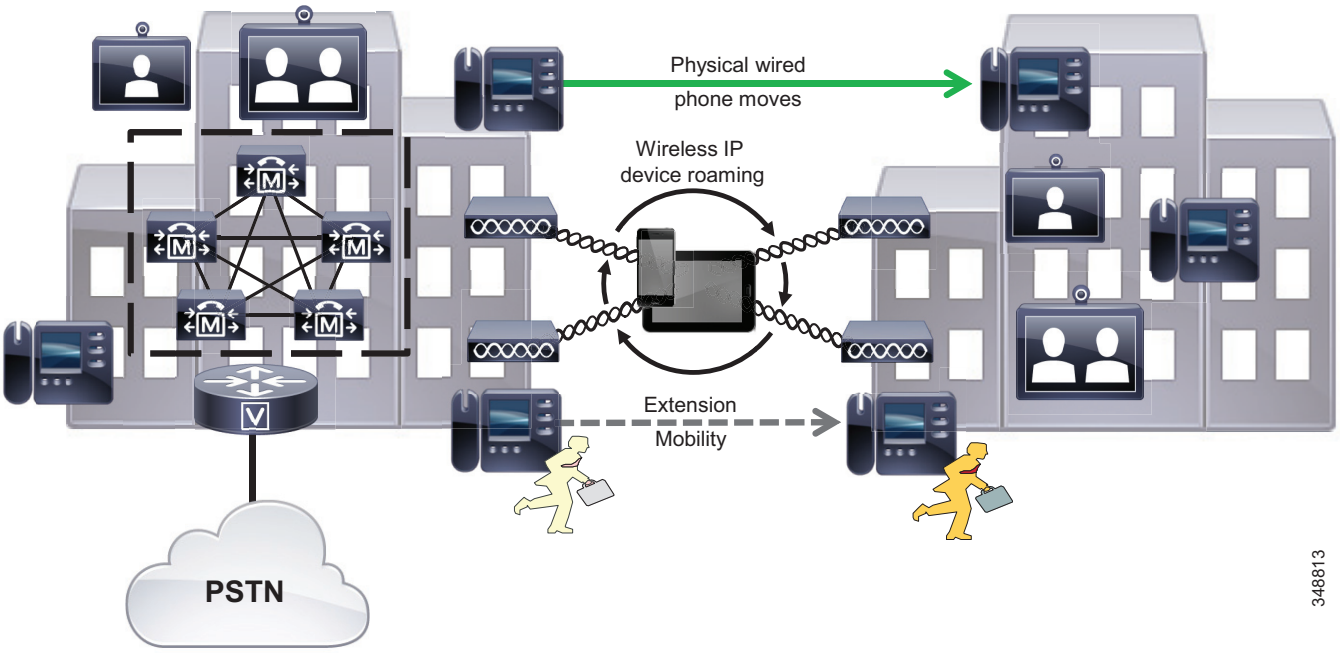
Campus Enterprise Mobility

Campus or single-site enterprise mobility refers to mobility within a single physical location typically bounded by a single IP address space and PSTN egress/ingress boundary. Mobility here not only includes the movement of users within this physical location but also the movement of endpoint devices.

Campus Enterprise Mobility Architecture

As illustrated in Figure 21-1, the enterprise campus mobility architecture is based on a single physical location that may include a single building or multiple buildings (as depicted) in close proximity, such that users are able to move freely within the campus and maintain IP and PSTN connectivity. Typically campus deployments involve a shared common connection or set of connections to the PSTN and Internet provider networks bound by a single IP address space and PSTN egress/ingress boundary. All users within this enterprise campus are connected to and reachable from a common network infrastructure.

Figure 21-1 Campus Enterprise Mobility Architecture



Types of Campus Mobility

Mobility within the campus enterprise typically involves the movement of devices, users, or both throughout the campus infrastructure. Campus enterprise mobility within Cisco Collaboration deployments can be divided into three main categories: physical wired phone movement, wireless device movement, and user movement without phone hardware or software. Each of these types of movements are discussed below.

Physical Wired Device Moves

As shown in Figure 21-1, movement of physical wired phones is easily accommodated within the campus infrastructure. These types of phone movements can occur within a single floor of a building, across multiple floors of a building, or even between buildings within the campus. Unlike with traditional PBX deployments where physical phone ports are fixed to a particular office, cubicle, or other space within the building, in IP telephony deployments a phone can be plugged into any IP port within the network infrastructure in order to connect to the IP PBX.

In a Cisco environment, this means a user can simply unplug a Cisco Unified IP Phone or Cisco TelePresence System endpoint from the network, pick it up and carry it to another location within the campus, and plug it into another wired network port. Once connected to the new network location, the phone simply re-registers to Unified CM and is able to make and receive calls just like in the previous location.

This same physical device movement also applies to software-based phones running on wired personal computers. For example, a user can move a laptop computer running Cisco IP Communicator or Cisco Jabber from one location to another within the campus, and after plugging the laptop into a network port in the new location, the software-based phone can re-register to Cisco call control and begin to handle phone calls again.

To accommodate physical device mobility within the campus, care should be taken when physically moving phone devices or computers running software-based phones to ensure that the network connection used at a new location has the same type of IP connectivity, connection speed, quality of service, security, and network services such as in-line power and dynamic host control protocol (DHCP), as were provided by the previous location. Failure to replicate these connection parameters, services, and features will lead to reduced functionality or in some cases complete loss of functionality.

Wireless Device Roaming

Wireless devices can move or roam throughout the enterprise campus, as shown in Figure 21-1, provided a wireless LAN network has been deployed to provide wireless network connectivity to the campus edge.

Examples of wireless devices include Cisco Unified Wireless IP Phones 7926G and 8821, wirelessly attached Cisco DX80, and Cisco mobile clients such as Cisco Jabber (see Cisco Mobile Clients and Devices, page 21-76).

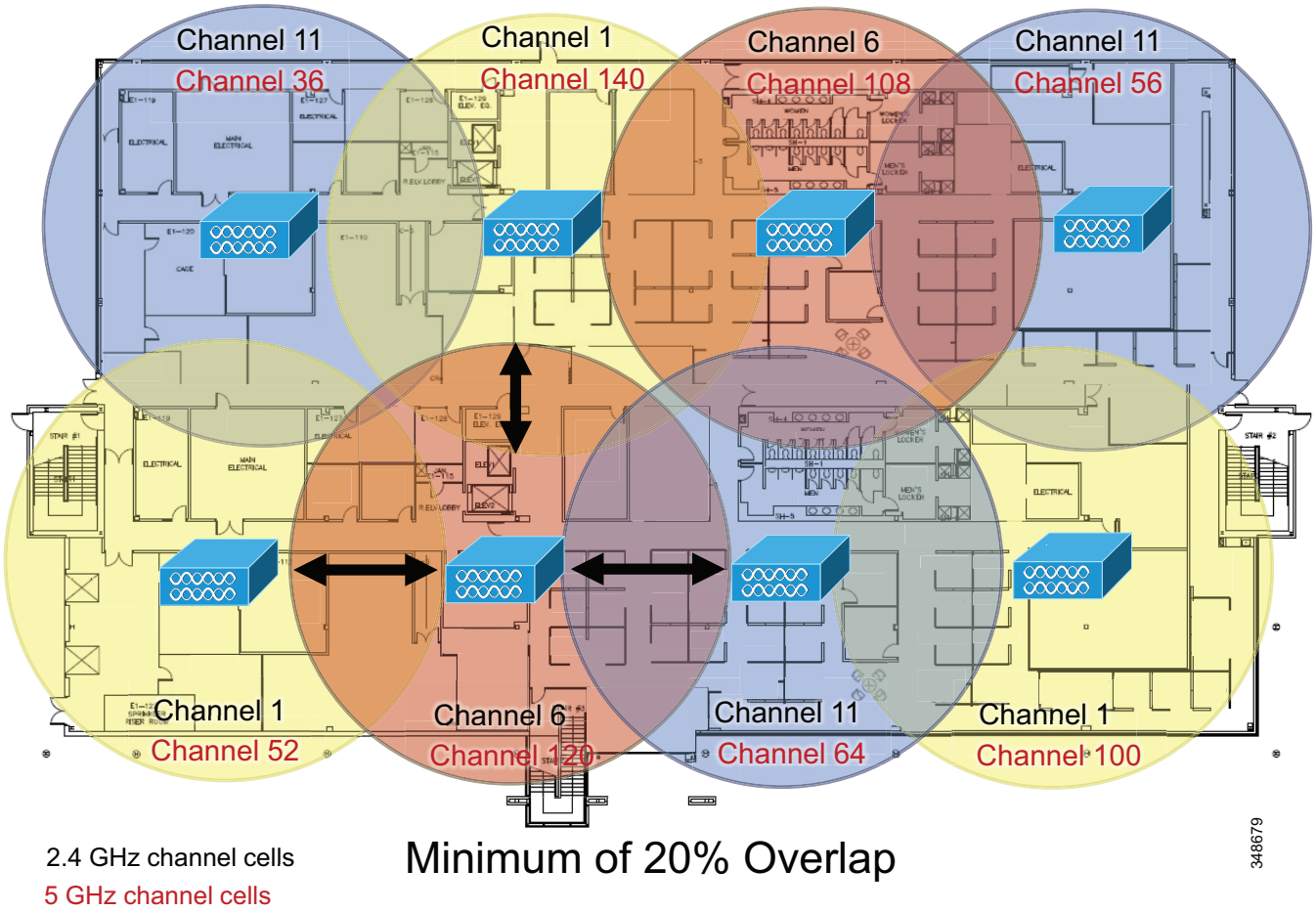
A WLAN network consists of one or more wireless access points (APs), which provide wireless network connectivity for wireless devices. Wireless APs are the demarcation point between the wireless network and the wired network. Multiple APs are deployed and distributed over a physical area of coverage in order to extend network coverage and capacity.

Because wireless devices and clients rely on the underlying WLAN infrastructure to carry both critical signaling and the real-time voice and video media traffic, it is necessary to deploy a WLAN network optimized for both data and real-time traffic. A poorly deployed WLAN network will be subjected to large amounts of interference and diminished capacity, leading not only to poor voice and video quality but in some cases dropped or missed calls. This will in turn render the WLAN deployment unusable for making and receiving voice calls. Therefore, when deploying wireless phones and clients, it is imperative to conduct a WLAN radio frequency (RF) site survey before, during, and after the deployment to determine appropriate cell boundaries, configuration and feature settings, capacity, and redundancy to ensure a successful voice and video over WLAN (VVoWLAN) deployment.

APs can be deployed autonomously within the network so that each AP is configured, managed, and operated independently from all other APs, or they can be deployed in a managed mode in which all APs are configured, managed, and controlled by a WLAN controller. In the latter mode, the WLAN controller is responsible for managing the APs as well as handling AP configuration and inter-AP roaming. In either case, to ensure successful VVoWLAN deployment, APs should be deployed using the following general guidelines:

- As shown in Figure 21-2, non-adjacent WLAN AP channel cells should overlap by a minimum of 20%. This overlap ensures that a wireless device can successfully roam from one AP to the next as the device moves around within the campus location while still maintaining voice and data network connectivity. A device that successfully roams between two APs is able to maintain an active voice call without any noticeable change in the voice quality or path.

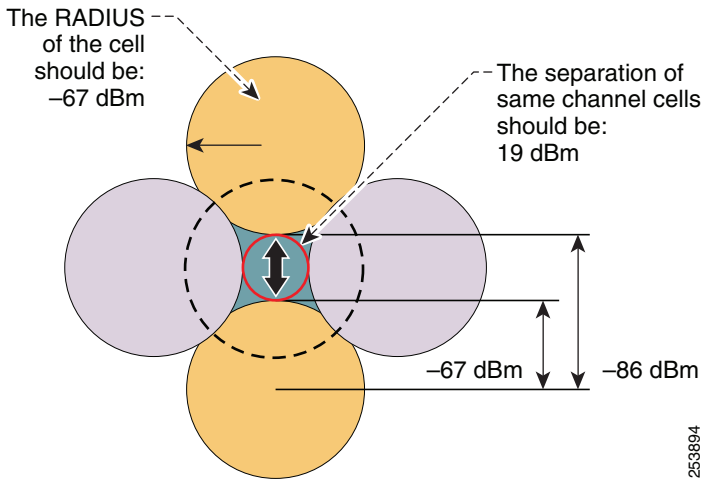
Figure 21-2 WLAN Channel Cell Overlap



- As shown in Figure 21-3, WLAN AP channel cells should be deployed with cell power-level boundaries (or channel cell radius) of -67 decibels per milliwatt (dBm). Additionally, the same-channel cell boundary separation should be approximately 19 dBm.

  A cell radius of approximately -67 dBm (or less) minimizes packet loss, which can be problematic for real-time voice and video traffic. A same-channel cell separation of 19 dBm is critical to ensure that APs or clients do not cause co-channel interference to other devices associated to the same channel, which would likely result in poor voice quality. The cell radius guideline of -67 dBm applies for both 2.4 GHz (802.11b/g/n) and 5 GHz (802.11a/n/ac) deployments.

Figure 21-3 WLAN Cell Radius and Same Channel Cell Separation



Note The 19 dBm same-channel cell separation is simplified and is considered ideal. It is very unlikely that this 19 dBm of separation can be achieved in most deployments. The most important RF design criteria are the -67 dBm cell radius and the minimum 20% recommended overlap between cells. Designing to these constraints optimizes channel separation.

Wireless roaming is not limited to wireless phones but also applies to software-based phones running on wireless personal computers. For example, a user can roam wirelessly throughout the campus with a laptop computer running Cisco IP Communicator or Cisco Jabber.

Most wireless APs, wireless phones, and wireless PC clients provide a variety of security options for providing secure access to the enterprise WLAN. In all cases, select a security method supported by both the WLAN infrastructure and the wireless devices that matches the security policies and requirements of the enterprise.

For more information on the Cisco Unified Wireless Network Infrastructure, see Wireless LAN Infrastructure, page 3-61. For more details on real-time traffic over WLAN design, including voice and video over WLAN, refer to the *Real-Time Traffic over Wireless LAN Solution Reference Network Design Guide*, available at

https://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Mobility/RToWLAN/CCVP_BK_R7805F20_00_rtowlan-srnd.html

Extension Mobility (EM)

As shown in Figure 21-1, in addition to physical movement of wired and wireless phones, the users themselves can also move around within the campus infrastructure without phone or PC hardware. In these cases, a user can move their enterprise extension or number from one device to another by applying a device profile containing the user's enterprise number and other settings.

The EM feature allows users to log on to IP phones located throughout the campus using a set of security credentials (user ID and PIN number). Once logged on, the user's personal device profile, including their enterprise phone number, calling privileges, and even their configured speed dials, is applied to the phone temporarily until the user logs out of the device or the login times out. The EM feature is available as part of Unified CM.

This feature is particularly useful for mobile enterprise users who spend considerable amounts of time outside the enterprise and are physically in the office only occasionally. By providing temporary office space for these types of mobile users, sometimes referred to as hot seating or free seating, a system administrator can accommodate large numbers of mobile users who only occasionally and temporarily need to use IP phone hardware.

To leverage EM within the campus the Unified CM administrator must configure user device profile(s) and user credentials, and subscribe IP phone(s) to the EM phone service.

Note EM is supported only with Unified CM call control and only on EM-capable endpoint devices.

For more information about EM, see Extension Mobility, page 18-7.

Campus Enterprise Mobility High Availability

Campus enterprise mobility features and solutions should be configured and deployed in a redundant fashion to ensure high availability of mobility functions and features.

For example, to effectively support hard-wired IP phones and computers running software-based IP phones, redundant and prevalent network connections or ports should be made available. Furthermore, these redundant network connections should be deployed with appropriate characteristics, including appropriate security, quality of service, and other network-based features to ensure optimal operation and voice quality for wired devices as they are moved from location to location. Ultimately a successful campus mobility deployment is possible only if the underlying network connectivity, PSTN connectivity, and other applications and services are deployed in a highly available fashion.

Likewise, when deploying or tuning a WLAN network for wireless device connectivity and roaming, it is also important to consider high availability for wireless services. To ensure resilient and sufficient coverage for the number of devices being deployed, a WLAN network should be deployed in a manner that ensures that adequate and redundant cells of coverage are provided without overlapping same-channel cells. Network connectivity for wireless devices and clients can be made highly available by providing ample cell coverage without same-channel cell overlap and sufficient overlap of different channel cells in order to facilitate roaming between APs.

Finally, when leveraging EM for user mobility within the campus, you should deploy this feature in a redundant fashion so that the failure of a single node within the Unified CM cluster does not prevent the operation of the Extension Mobility feature. For information on deploying Cisco Extension Mobility in a highly available manner, see High Availability for Extension Mobility, page 18-15.

Capacity Planning for Campus Enterprise Mobility

Deploying campus enterprise mobility successfully requires providing ample capacity to accommodate all mobile users exercising these mobility features and solutions.

Capacity considerations for physical movement of wired devices and computers depend completely on the number of network ports that are made available within the campus network infrastructure. In order for users to move devices around the campus, there must be some number of available network ports in each location that can be used to connect these mobile users' devices. A shortage of network ports to accommodate this wired device movement can result in an inability to move a device physically from one location to another.

When deploying wireless devices and leveraging wireless device roaming within the enterprise WLAN, it is also important to consider the device connectivity and call capacity of the WLAN infrastructure. Oversubscription of the campus WLAN infrastructure in terms of number of devices or number of active calls will result in dropped wireless connections, poor voice and video quality, and delayed or failed call setup. The chances of oversubscribing a deployment of voice and video over WLAN (VVoWLAN) are greatly minimized by deploying sufficient numbers of APs to handle required call capacities. AP call capacities are based on the number of simultaneous voice and/or video bidirectional streams that can be supported in a single channel cell area. The general rule for VVoWLAN call capacities is as follows:

- Maximum of 27 simultaneous voice over WLAN (VoWLAN) bidirectional streams per 802.11g/n (2.4 GHz) channel cell with Bluetooth disabled and 24 Mbps or higher data rates.
- Maximum of 27 simultaneous VoWLAN bidirectional streams per 802.11a/n/ac (5 GHz) channel cell with 24 Mbps or higher data rates.
- Assuming a video resolution of 720p (high-definition) and a video bit rate of up to 1 Mbps, a maximum of 8 simultaneous VVoWLAN bidirectional streams per 802.11 g/n (2.4 GHz) with Bluetooth disabled or 802.11 a/n/ac (5 GHz) channel cell.

These voice and video call capacity values are highly dependent upon the RF environment, the configured or supported video resolution and bit rates, the wireless endpoint and its specific capabilities, and the underlying WLAN system features. Actual capacities for a particular deployment could be less.

Note A single call between two wireless endpoints associated to the same AP is considered to be two simultaneous bidirectional streams.

Scalability of EM is dependent almost completely on the login/logout rate of the feature within Unified CM. It is important to know the number of extension mobility users enabled within the Unified CM cluster as well as how many users are moving around the campus and exercising this feature at any given time to ensure that sufficient EM login/logout capacity can be provided to these mobile users. For more information on EM capacity planning, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

In all cases, the Unified CM cluster(s) within the campus must have sufficient device registration capacity to handle device registration for moved devices, regardless of whether they are wired or wireless devices. Of course, assuming all devices being moved throughout the campus are already deployed within the campus network, then sufficient capacity within the call control platform should already be in place prior to the movement of devices. If new devices are added to the deployment for mobility purposes, however, device registration capacity should be considered and, if necessary, additional capacity should be added.

Finally, given the many features and functions provided by Unified CM, configuration and deployment of these mobility solutions does have sizing implications for the overall system. Determining actual system capacity is based on considerations such as number of endpoint devices, EM users, and busy hour call attempt (BHCA) rates to number of CTI applications deployed. For more information on general system sizing, capacity planning, and deployment considerations, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Design Considerations for Campus Enterprise Mobility

Observe the following design recommendations when deploying campus enterprise mobility features and solutions:

- To accommodate physical device mobility within the campus ensure that the network connection used at a new location has the same type of IP connectivity (VLANs, inter-VLAN routing, and so forth), connection speed, quality of service, security, and network services (in-line power, dynamic host control protocol (DHCP), and so forth) as provided by the previous network connection. Failure to replicate these connection parameters, services, and features will lead to diminished functionality and in some case complete loss of functionality.
- When deploying wireless IP devices and software-based clients, it is imperative to conduct a WLAN radio frequency (RF) site survey before, during, and periodically after the deployment to determine appropriate cell boundaries, configuration and feature settings, capacity, and redundancy to ensure a successful voice and video over WLAN (VVoWLAN) deployment.
- APs should be deployed with a minimum cell overlap of 20%. This overlap ensures that a dual-mode device can successfully roam from one AP to the next as the device moves around within a location, while still maintaining voice and data network connectivity.
- APs should be deployed with cell power level boundaries (or channel cell radius) of -67 dBm in order to minimize packet loss. Furthermore, the same-channel cell boundary separation should be approximately 19 dBm. A same-channel cell separation of 19 dBm is critical for ensuring that APs or clients do not cause co-channel interference to other devices associated to the same channel, which would likely result in poor voice and video quality.
- Deploy EM services in a highly redundant manner so that the loss of a single Unified CM node does not have adverse effects on the feature operation. If EM services are critical, consider deploying a server load balancing solution to route around Unified CM node failures and provide highly available functionality. For more information on EM high availability, see High Availability for Extension Mobility, page 18-15.
- Provide sufficient wireless voice and video call capacity on the campus network by deploying the appropriate number of wireless APs to handle the desired call capacity based on wireless user BHCA rates. Each 802.11g/n (2.4 GHz) or 802.11a/n/ac (5 GHz) channel cell can support a maximum of 27 simultaneous voice-only calls with 24 Mbps or higher data rates. Each 802.11g/n (2.4 GHz) or 802.11a/n/ac (5 GHz) channel cell can support a maximum of 8 simultaneous video calls assuming 720p video resolution at up to 1 Mbps bit rate. For 2.4 GHz WLAN deployments, Bluetooth must be disabled to achieve this capacity. Actual call capacity could be lower depending on RF environment, wireless endpoint type, and WLAN infrastructure.

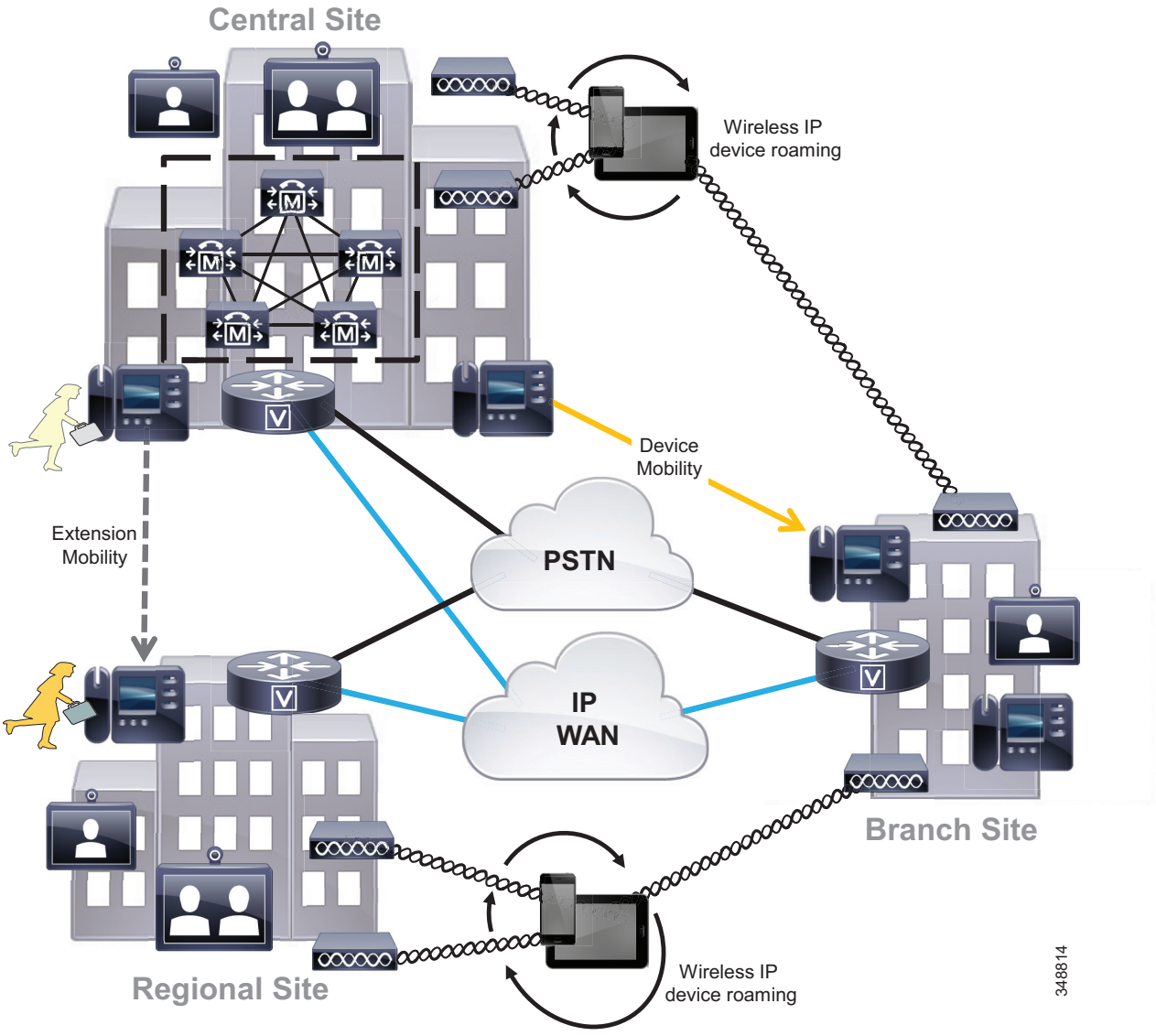
Multisite Enterprise Mobility

Multisite enterprise mobility refers to mobility within an enterprise with multiple physical locations, each with a unique IP address space and PSTN egress/ingress boundary. Mobility in this case includes not only the movement of users and endpoint devices within each physical location but also movement of users and endpoint devices between sites and locations.

Multisite Enterprise Mobility Architecture

As shown in Figure 21-4, the multisite enterprise mobility architecture is based on two or more locations or sites geographically separated. Sites may vary in size from large numbers of users and devices in a central or campus site to smaller numbers of users and devices in medium-sized regional sites or smaller branch sites. Typically multisite enterprise deployments consist of IP WAN links interconnecting sites as well as local PSTN egress/ingress at each location. In addition, critical services are often replicated at each physical site in order to maintain features and functions during network outages between sites. From a mobility perspective, users and their devices may be mobile within a site or between sites.

Figure 21-4 Multisite Enterprise Mobility Architecture



Note While Figure 21-4 depicts a multisite deployment with centralized call processing (as evidenced by a single Unified CM cluster within the central site), the same design and deployment considerations for multisite enterprise mobility deployments apply to distributed call processing environments. Differences in mobility feature operation when deployed in distributed call processing environments are described in the following discussions.

Types of Multisite Enterprise Mobility

Mobility within a multisite enterprise deployment involves not only the movement of devices, users, or both within a single site, but also movement of users and devices between sites.

The same types of mobility features and solutions supported with campus or single site enterprise deployments apply to intra-site movement of users and devices within any single site of a multisite deployment. These include physical wired phone movement, wireless phone roaming, and extension mobility. For information on these types of mobility solutions and functions, see Campus Enterprise Mobility, page 21-4.

For inter-site mobility in a multisite deployment, these same mobility features are also supported in much the same way. However, the key difference with these features when applied between two or more sites is that they are augmented with the Device Mobility feature. The Device Mobility feature provides a mechanism for dynamic location awareness of devices based on the IP address the device uses when connecting to the enterprise network.

Physical Wired Device Moves

Movement of physical wired phones is easily accommodated within each site of a multisite deployment as well as between sites. Just as with a campus or single-site deployment, wired device movement limited to a single site of a multisite deployment simply involves unplugging a Cisco endpoint from the network, moving it to another location within the site, and plugging it into another wired network port. Once connected to the new network location, the phone simply re-registers to the call control platform and is able to make and receive calls just like in the previous location.

Movement of wired devices between sites or locations in a multisite deployment involve the same basic behavior. However, the Device Mobility feature, when combined with this type of mobility, ensures that call admission control operations and gateway and codec selection are appropriate once the device re-registers in the new location to which it has been moved. See Device Mobility, page 21-14, for information about this feature.

Wireless Device Roaming

Just as with a single-site campus deployment, wireless devices can move or roam throughout a multisite enterprise deployment, as shown in Figure 21-4, provided wireless LAN network infrastructure is available at each site to provide wireless network connectivity. However, as with the movement of wired phones between sites, the Device Mobility feature should also be deployed for wireless devices to ensure that the correct gateway and codec are used when making and receiving calls and that call admission control manages bandwidth appropriately. See Device Mobility, page 21-14, for information about this feature.

For distributed call processing environments, just as with wired phones, wireless devices should be configured to register with only a single call processing platform or cluster to avoid potential issues with call routing.

Extension Mobility (EM)

In addition to supporting EM within a single site, as illustrated in Figure 21-4, this feature is also supported between sites to enable users to move between sites within the enterprise and log on to phones in each locations.

EM is also supported in distributed call processing deployments when users move between sites and phones on different Unified CM clusters. To support extension mobility in distributed call processing environments, you might need to configure the Cisco Extension Mobility Cross Cluster (EMCC) feature. For information about this feature, see Extension Mobility Cross Cluster (EMCC), page 18-9.

Note EM and EMCC are supported only with Unified CM call control and only on EM-capable endpoint devices.

Device Mobility

With Cisco Unified CM, a site or a physical location is identified using various settings such as locations, regions, calling search spaces, and media resources. Cisco Unified IP Phones residing in a particular site are statically configured with these settings. Unified CM uses these settings for proper call establishment, call routing, media resource selection, and so forth. However, when dual-mode phones and other mobile client devices such as Cisco Unified Wireless IP Phones are moved from their home site to a remote site, they retain the home settings that are statically configured on the phones. Unified CM then uses these home settings on the phones in the remote site. This situation is undesirable because it can cause problems with call routing, codec selection, media resource selection, and other call processing functions.

Cisco Unified CM uses a feature called Device Mobility, which enables Unified CM to determine if the IP phone is at its home location or at a roaming location. Unified CM uses the device's IP subnets to determine the exact location of the IP phone. By enabling device mobility within a cluster, mobile users can roam from one site to another, thus acquiring the site-specific settings. Unified CM then uses these dynamically allocated settings for call routing, codec selection, media resource selection, and so forth.

This section begins with a discussion surrounding the main purpose for the Device Mobility feature, followed by an in-depth discussion of the Device Mobility feature itself. This discussion covers the various components and configuration constructs of the Device Mobility feature. This section also presents an in-depth discussion of the impact of the Device Mobility feature on the enterprise dial plan, including the implication for various dial plan models.

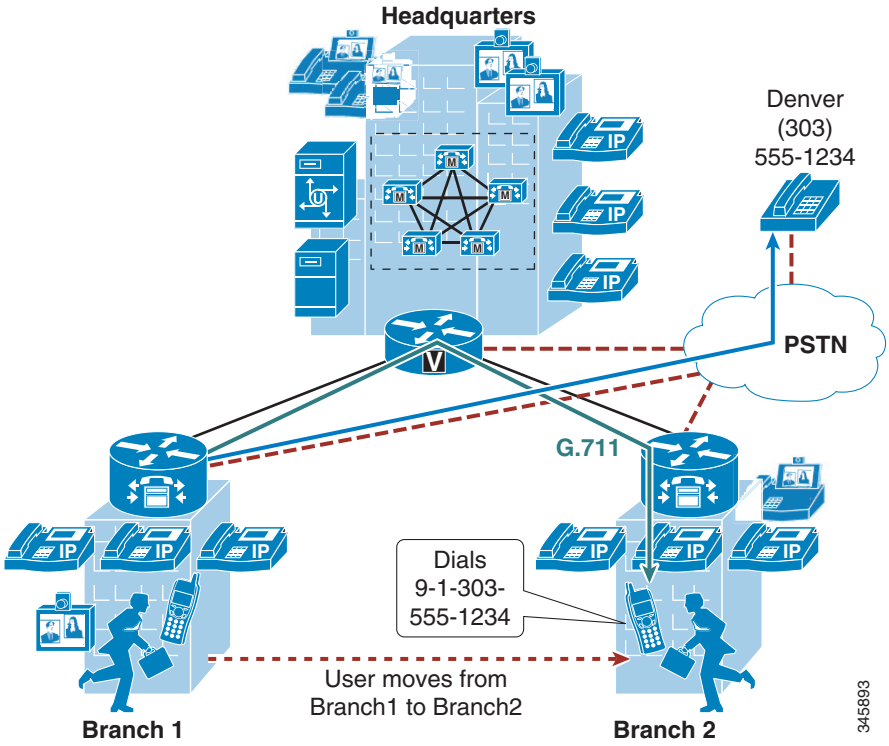
Note Device mobility is supported only with Unified CM call control.

Need for Device Mobility

This section explains the need for device mobility when there are many mobile users in a Unified CM cluster.

Figure 21-5 illustrates a hypothetical network containing a Unified CM cluster without the Device Mobility feature, located at the headquarter site (HQ). The cluster has two remote sites, Branch1 and Branch2. All intra-site calls use G.711 voice codecs, while all inter-site calls (calls across the IP WAN) use G.729 voice codecs. Each site has a PSTN gateway for external calls.

Figure 21-5 Example Network with Two Remote Sites



When a user in Branch1 moves to Branch2 and calls a PSTN user in Denver, the following behavior occurs:

- Unified CM is not aware that the user has moved from Branch1 to Branch2. An external call to the PSTN is sent over the WAN to the Branch1 gateway and then out to the PSTN. Thus, the mobile user continues to use its home gateway for all PSTN calls.
- The mobile user and Branch1 gateway are in the same Unified CM region and location. Location-based call admission control is applicable only for devices in different locations, and an intra-region call uses the G.711 voice codec. Thus, the call over the IP WAN to the Branch1 gateway uses the G.711 codec and is not tracked by Unified CM for purposes of call admission control. This behavior can result in over-subscription of the IP WAN bandwidth if all the remote links are low-speed links.
- The mobile user creates a conference by adding multiple Branch2 users to the existing call with the PSTN user in Denver. The mobile user uses the conferencing resource that is on the Branch1 gateway, therefore all conference streams flow over the IP WAN.

Note Device Mobility is an intra-cluster feature and does not span multiple Unified CM clusters. In distributed call processing environments, Device Mobility must be enabled and configured on each Unified CM cluster within the deployment.

Note In deployments where Device Mobility is not configured, administrators may wish to over-provision WAN bandwidth between site locations to ensure that physical movement of devices across the WAN and between sites does not over-subscribe the WAN. The amount of bandwidth to over-provision on each WAN link depends on the anticipated rate at which users will move devices between two locations.

Device Mobility Architecture

The Unified CM Device Mobility feature helps solve the problems mentioned above. This section briefly explains how the feature works. However, for a detailed explanation of this feature, refer to the Device Mobility information in the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

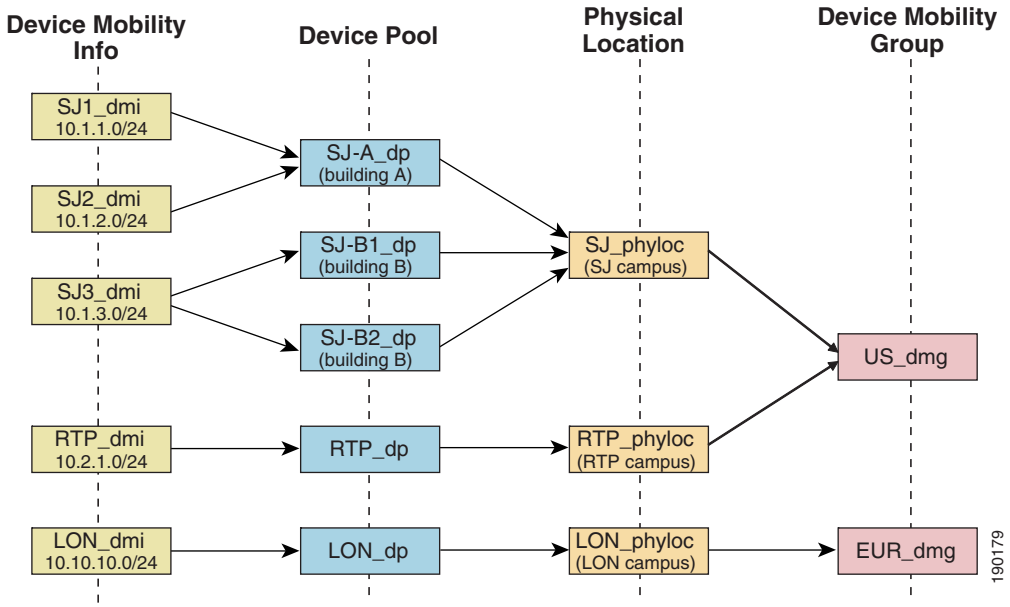
https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

Some of the device mobility elements include:

- Device Mobility Info — Configures IP subnets and associates device pools to the IP subnets.
- Device Mobility Group — Defines a logical group of sites with similar dialing patterns (for example, US_dmg and EUR_dmg in Figure 21-6).
- Physical Location — Defines the physical location of a device pool. In other words, this element defines the geographic location of IP phones and other devices associated with the device pool. (For example, all San Jose IP phones in Figure 21-6 are defined by physical location SJ_phyloc.)

Figure 21-6 illustrates the relationship between all these terms.

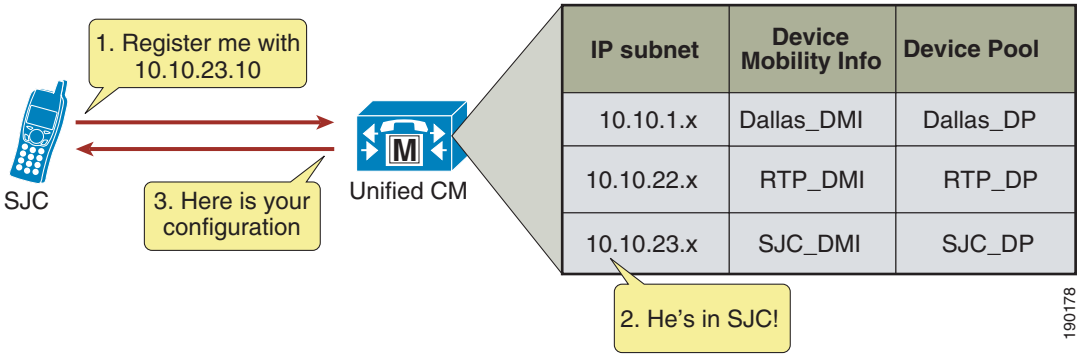
Figure 21-6 Relationship of Device Mobility Components



Unified CM assigns a device pool to an IP phone based on the device's IP subnet. The following steps, illustrated in Figure 21-7, describe the behavior:

1. The IP phone tries to register to Unified CM by sending its IP address in the Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) registration message.
2. Unified CM derives the device's IP subnet and matches it with the subnet configured in the Device Mobility Info.
3. If the subnet matches, Unified CM provides the device with a new configuration based on the device pool configuration.

Figure 21-7 Phone Registration Process



Unified CM uses a set of parameters under the device pool configuration to accommodate Device Mobility. These parameters are of the following two main types:

- Roaming Sensitive Settings, page 21-18
- Device Mobility Related Settings, page 21-19

Roaming Sensitive Settings

The parameters under these settings will override the device-level settings when the device is roaming within or outside a Device Mobility Group. The parameters included in these settings are:

- Date/time Group
- Region
- Media Resource Group List
- Location
- Network Locale
- SRST Reference
- Physical Location
- Device Mobility Group

The roaming sensitive settings primarily help in achieving proper call admission control and voice codec selection because the location and region configurations are used based on the device's roaming device pool.

For more details on various call admission control techniques, see the chapter on Bandwidth Management, page 13-1.

The roaming sensitive settings also update the media resource group list (MRGL) so that appropriate remote media resources are used for music on hold, conferencing, transcoding, and so forth, thus utilizing the network efficiently.

The roaming sensitive settings also update the Survivable Remote Site Telephony (SRST) gateway. Mobile users register to a different SRST gateway while roaming. This registration can affect the dialing behavior when the roaming phones are in SRST mode.

For example, if a user moves with their phone to a new location that loses connectivity to Unified CM, then based on the roaming sensitive Device Mobility settings, a new SRST reference is configured for the moved phone and the moved phone will now be under control of the local roaming location SRST router. When this occurs, not only would the user's phone be unreachable from the PSTN or other sites because the device’s DID will not have changed and will still be anchored at their home location, but in addition reachabililty from devices within the local failed site might be difficult without the use of abbreviated dialing as implemented within SRST.

As an example, assume that a user moves a phone from their home location in San Jose, which has a directory number of 51234 and an associated DID of 408 555 1234 to a remote location in New York, and that the link between the New York site and San Jose fails shortly after the user roams to the New York location. In this scenario the phones in the New York site will all fail-over to the SRST router in that site. The roaming/moved phone will also register to the New York SRST router because its SRST reference was updated based on the device mobility roaming sensitive settings. In this scenario, the local New York devices will register to the SRST router with five-digit extensions just as they do to Unified CM, and as a result the roaming phone still has a directory number of 51234. To reach the roaming phone from all other sites and from the PSTN, the number 408 555 1234 will be routed to the San Jose PSTN gateway to which this particular DID is anchored. Because the New York site is disconnected from the San Jose site, any such calls will be routed to the users’ voicemail boxes since they will be unreachable at their desk phones. Likewise, calls internally within the local failed site will

have to be dialed using five-digit abbreviated dialing or based on the configured digit prefixing as defined by the **dialplan-pattern** and **extension-length** commands within the SRST router. In either case, local callers will have to be understand the required dialing behavior for reaching the local roaming device by abbreviated dialing. In some cases this may be simply five-digit dialing or it may be that users have to dial a special digit prefix to reach the local roaming phone. The same logic applies to outbound dialing from the moved or roaming phone in New York because its dialing behavior might have to be altered in order to reach local extensions using abbreviated dialing. Outbound dialing to the PSTN from the local roaming device should remain the same, however.

Device Mobility Related Settings

The parameters under these settings will override the device-level settings only when the device is roaming within a Device Mobility Group. The parameters included in these settings are:

- Device Mobility Calling Search Space
- AAR Calling Search Space
- AAR Group
- Calling Party Transformation CSS

The device mobility related settings affect the dial plan because the calling search space dictates the patterns that can be dialed or the devices that can be reached.

Device Mobility Group

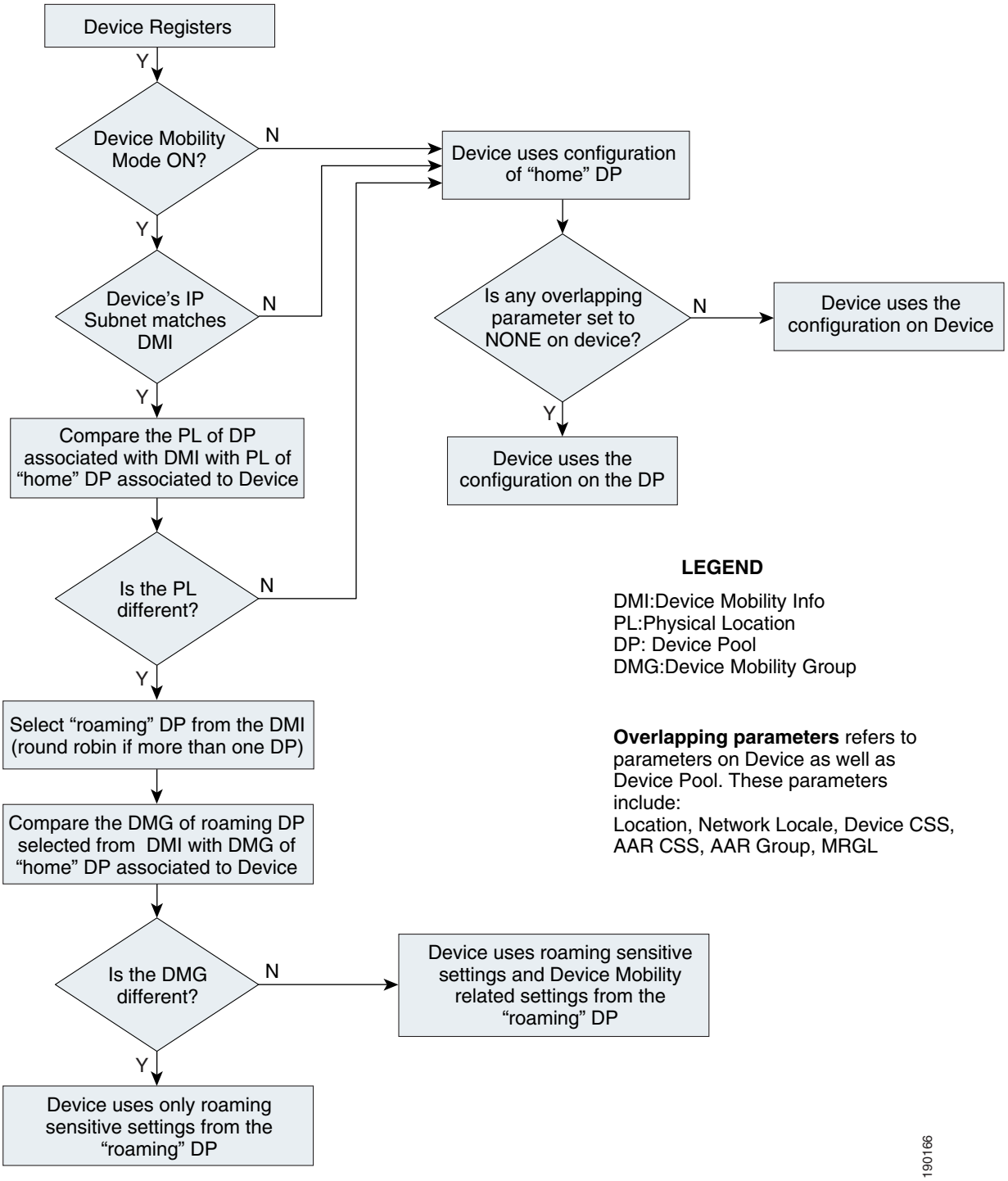
Device Mobility Group, as explained earlier, defines a logical group of sites with similar dialing patterns (for example, sites having the same PSTN access codes and so forth). With this guideline, all sites have similar dialing patterns in the site-specific calling search spaces. Sites having different dialing behavior are in a different Device Mobility Group. As illustrated in Figure 21-6, the San Jose and RTP sites' Device Mobility Info, Device Pools, and Physical Locations are different; however, all of these have been assigned to the same Device Mobility Group US_dmg because the required dialing patterns and PSTN access codes are the same between the two locations. On the other hand, the London site is assigned to a separate Device Mobility Group EUR_dmg due to the fact that the required dialing patterns and PSTN access codes there are different than those of the US sites. A user roaming within a Device Mobility Group may preserve his dialing behavior at the remote location even after receiving a new calling search space. A user roaming outside the Device Mobility Group may still preserve his dialing behavior at the remote location because he uses his home calling search space.

However, if a Device Mobility Group is defined with sites having different dialing patterns (for example, one site requires users to dial 9 to get an outside line while another site requires users to dial 8 to get an outside line), then a user roaming within that Device Mobility Group might not preserve his same dialing behavior at all locations. A user might have to dial digits differently at different locations after receiving a new calling search space at each location. This behavior can be confusing for users, therefore Cisco recommends against assigning sites with different dialing patterns to the same Device Mobility Group.

Device Mobility Operation

The flowchart in Figure 21-8 represents the operation of the Device Mobility feature.

Figure 21-8 Operation of the Device Mobility Feature



The following guidelines apply to the Device Mobility feature:

- If the overlapping parameters listed in Figure 21-8 have the same configurations on the device as well as the device pool, then these parameters may be set to NONE on the device. These parameters must then be configured on the device pool. This practice can greatly reduce the amount of configuration because the devices do not have to be configured individually with all the parameters.
- Define one physical location per site. A site may have more than one device pool.
- Define sites with similar dialing patterns for PSTN or external/off-net access with the same Device Mobility Group.
- A "catch-all" Device Mobility Info with IP subnet 0.0.0.0 may be defined for all non-defined subnets, depending on the company policy. This Device Mobility Info may be used to assign a device pool that can restrict access or usage of the network resources. (For example, the device pool may be configured with a calling search space NONE that will block any calls from the device associated with this device pool while roaming.) However, by doing so, administrators must be aware of the fact that this will block all calls, even 911 or other emergency calls. The calling search space may be configured with partitions that will give access only to 911 or other emergency calls.

Dial Plan Design Considerations

The Device Mobility feature uses several device and device pool settings that are based on the settings in the roaming device pool selected and on the IP address with which the endpoint registers. For details of which settings are updated with the settings of the device pool for the subnet, refer to the Device Mobility information in the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

From the dial plan perspective, mainly the AAR group, AAR CSS, device CSS, Local Route Group, and outgoing call's calling party transformation CSS settings are relevant.

Egress Gateway Selection for Roaming Devices

Typically the desired egress gateway selection behavior of roaming devices is to use gateways local to the visited site. The recommended way to implement egress gateway selection that is specific to the calling device is to use PSTN route patterns pointing to route lists that use Standard Local Route Group. Using Standard Local Route Group in a route list effectively means that Standard Local Route Group, when routing an actual call, will be replaced with the Local Route Group configured in the device pool of the calling endpoint. This schema ensures that site-unspecific route patterns and route lists are used; site-specific egress gateway selection completely relies on device pool-level Local Route Group configuration.

For roaming devices (whether roaming inside or between device mobility groups), the device mobility feature always ensures that the Local Route Group of the roaming device pool is used as Standard Local Route Group. This guarantees that, with Local Route Group egress gateway selection, a visited site-specific route group (and thus gateways local to the visited site) will typically be used. This behavior ensures that, for example, emergency calls routed via route patterns that use a Standard Local Route Group route list will always use egress gateways local to the visited site.

Local Route Group egress gateway selection can be used with all dial plan approaches explained in the chapter on Dial Plan, page 14-1.

If certain calls from roaming endpoints need to be routed through gateways local to the home site of the roaming phone, then routing for these calls has to be implemented through route patterns pointing to route lists that use fixed site-specific route groups instead of Standard Local Group.

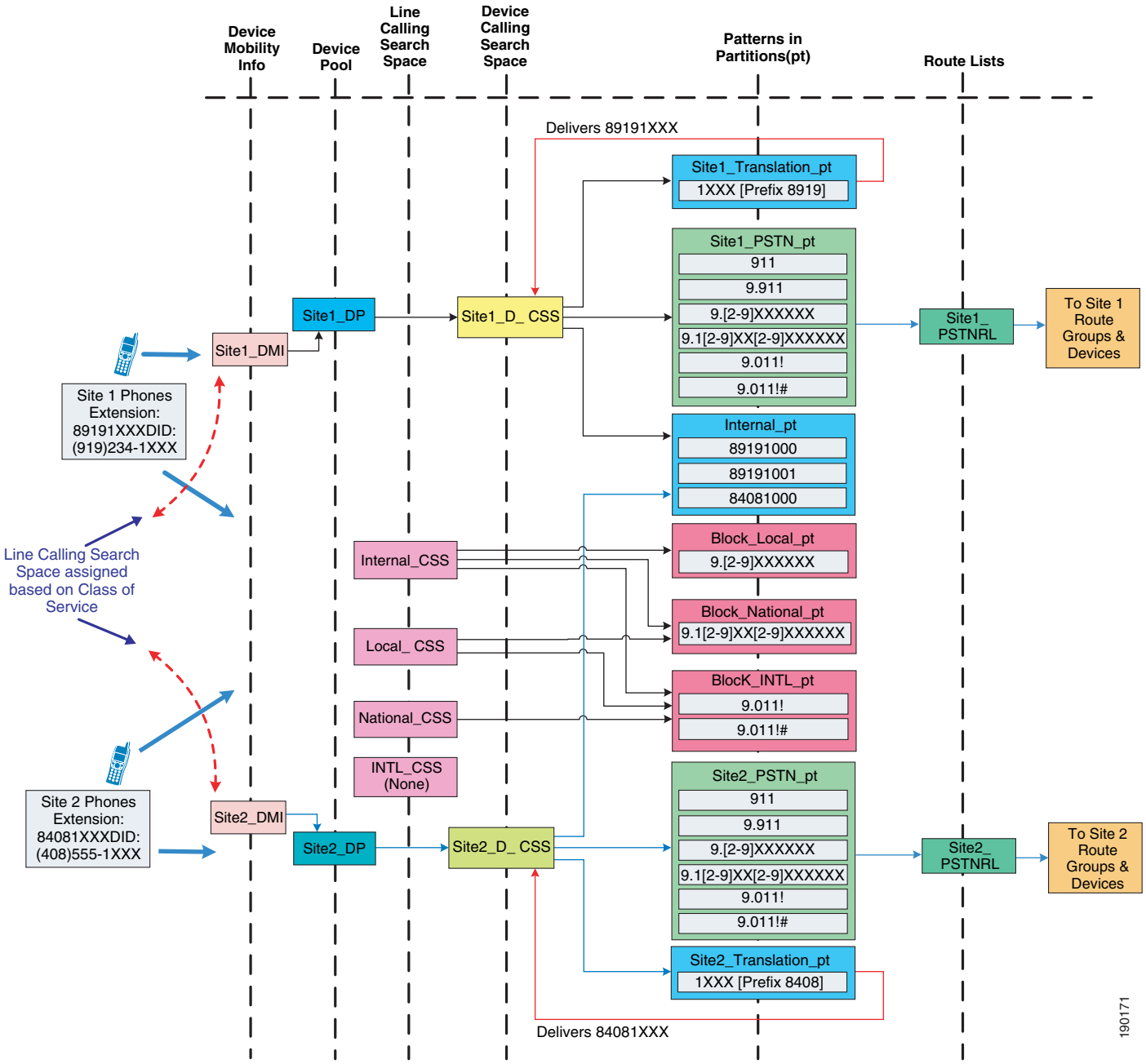
In a line/device dial plan approach, these route patterns would be addressed by the device CSS configured on the endpoint. When roaming but not leaving the device mobility group, the calling endpoint's device CSS is replaced by the Device Mobility CSS configured on the roaming device pool. If fixed egress gateway selection is required for some calls and the route patterns for those calls are addressed by the device CSS, you have to make sure that roaming devices always roam across device mobility groups. This will guarantee that roaming endpoints always use the device CSS configure on the endpoint.

When using the +E.164 dial plan approach explained in the chapter on Dial Plan, page 14-1, all PSTN route patterns are accessible by the line CSS, which is not changed or updated for roaming devices. In this dial plan, site-specific route patterns tying specific PSTN destinations to fixed gateways (for example, in the home location of the roaming device) are not affected by device mobility operation.

Variable Length On-Net Dialing with Flat Addressing Using the Line/Device Approach without Local Route Group

Figure 21-9 shows a variable-length on-net dial plan with flat addressing for Device Mobility.

Figure 21-9 Variable-Length On-Net Dial Plan with Flat Addressing for Device Mobility



The following design considerations apply to the dial plan model in Figure 21-9:

- In this dial plan the translation patterns implementing 4-digit intra-site dialing are addressed by the device CSS. This is done to avoid the requirement to have site-specific line CSSs. Mobile users inherit the intra-site dialing of the visited site because the device CSS is updated with the roaming device pool's device mobility CSS (assuming the user is roaming inside the device mobility group). If this behavior is not desired, consider defining each site as a Device Mobility Group. However, users must be aware that, for any external PSTN calls, the mobile phone continues to use the home gateway and therefore consumes WAN bandwidth. This can be avoided by using Standard Local Route Group (see Egress Gateway Selection for Roaming Devices, page 21-21).
- Additional device calling search spaces may be configured for roaming users with access only to the PSTN and internal phones partitions. This configuration will need at least one additional device pool and calling search space per site. Thus, *N* sites will need *N* device pools and *N* calling search spaces. However, this configuration will not require defining each site as a Device Mobility Group. With this configuration mobile users, when roaming, will not have access to dialing habits through translation patterns in their device CSS.
- Mobile users registered with a remote SRST gateway have unique extensions. However, mobile users must be aware that no PSTN user can call them when they are registered to a remote SRST gateway.

+E.164 Dial Plan with Traditional Approach and Local Route Group

As described in the chapter on Dial Plan, page 14-1, the line/device approach has some specific issues, and creating a +E-164 dial plan based on the line/device approach is not recommended. The recommended approach for +E.164 dial plans is to combine class of service selection and dialing normalization on the line CSS and use the Local Route Group feature to address the requirement for site-specific egress gateway selection. In this approach the device CSS on the phone is not used at all. If you combine this approach with device mobility, the only roaming sensitive component of the design is the device pools' local route group. For a roaming phone (whether roaming inside or between device mobility groups), the local route group defined on the phone's home device pool will always be updated with the local route group defined on the roaming device pool. This guarantees that all calls always egress through a gateway local to the visited site.

Multisite Enterprise Mobility High Availability

Multisite enterprise mobility features and solutions should be configured and deployed in a redundant fashion in order to ensure high availability of mobility functionality. High availability considerations for wired phone moves, wireless roaming, and EM in multisite mobility deployments are similar to those for campus mobility deployments. Just as with campus environments, redundant network ports, wireless cell coverage, and Unified CM nodes handling extension mobility logins and logouts should be provided to ensure highly available services.

Similarly, it is important to consider high availability of the Device Mobility feature. Because Device Mobility is natively integrated within Unified CM call control, the failure of a cluster node should have no impact on the functionality of Device Mobility. Device pool, Device Mobility Info, Device Mobility Group, and all other configurations surrounding Device Mobility are preserved if there is a failure of the publisher node or a call processing (subscriber) node. Additionally, if there is a call processing node failure, affected phones will fail-over to their secondary call processing node or Survivable Remote Site Telephony (SRST) reference router as usual based on the Unified CM Group construct.

Note Cisco TelePresence System endpoints do not support registration redundancy with Cisco IOS SRST.

Capacity Planning for Multisite Enterprise Mobility

As for Device Mobility scalability considerations, there are no specific or enforced capacity limits surrounding this feature and the various configuration constructs (device pools, device mobility groups, and so forth). For more information on general system sizing, capacity planning, and deployment considerations, see the chapter on Collaboration Solution Sizing Guidance, page 25-1

Design Considerations for Multisite Enterprise Mobility

All campus enterprise mobility design considerations apply to multisite enterprise mobility deployments as well (seeDesign Considerations for Campus Enterprise Mobility, page 21-11). The following additional design recommendations apply specifically to multisite mobility environments:

- Ensure that all critical services (device registration, PSTN connectivity, DNS, DHCP, and so forth) are deployed at each site in a multisite deployment so that failure of the connection between the site and other sites does not disrupt critical operations. In addition, ensure that a sufficient number of physical network ports and wireless LAN APs are available at each site to support movement of devices and required call capacity.
- In situations in which sites with different dialing patterns (for example, sites having different PSTN access codes) are configured in the same Device Mobility Group, roaming users might have to dial numbers differently based on their location, which can be confusing. For this reason, Cisco recommends assigning sites with similar dialing patterns (for example, sites having the same PSTN access codes) to the same Device Mobility Group. Doing so ensures that roaming users can dial numbers the same way at all sites within the Device Mobility Group.
- The Device Mobility settings from the "roaming" device pool are applied only when users roam within the same Device Mobility Group; therefore, avoid roaming between different Device Mobility Groups because the resulting call routing behavior will cause originated calls from the moved phone to be routed using the "home" or device-configured calling search space. This can lead to unnecessary consumption of WAN bandwidth because the call might be routed through a different site's gateway rather than the local "roaming" gateway.
- Define only one physical location per site. This ensures that device mobility is engaged only in scenarios in which a user is roaming between sites. For roaming within the same site, the concerns that mandate Device Mobility (for example, WAN bandwidth consumption, codec selection, and call admission control) are not present because low-speed links typically are not deployed within a single site.
- In failover scenarios, "roaming" phones will utilize the SRST reference/gateway as dictated by the "roaming" device pool's roaming sensitive settings. Therefore, in these situations the "roaming" phone is unreachable from the PSTN due to the fact that the DID for this phone is anchored in another location's PSTN gateway. Furthermore, for outbound calls from the "roaming" phone, dialing behavior might have to be altered for things such as PSTN access codes, and speed dials configured on the phone might not be usable.
- If your system requires the ability to use abbreviated dialing or to use speed dials that rely on abbreviated dialing, Cisco recommends using a Uniform On-net dial plan model because it will ensure that abbreviated dialing (direct or through speed dials) continues to work even when the mobile user's phone is in a roaming location. Abbreviated dialing is still possible with this dial plan model because all extensions or directory numbers are unique across all sites, and therefore abbreviated dialing can be used universally due to the fact that there are no overlapping extensions.

- If your system uses a Variable Length On-net dial plan model (using either the line/device or the line-CSS-only +E.164 dial plan approach), Cisco recommends configuring speed dials in a universal way so that a single unique extension can be reached when called. By configuring speed dials using full +E.164 numbers or using site or access codes, you can enable roaming users to use the same speed dials at any location.
- If Device Mobility is enabled for users who on occasion access the enterprise network through a VPN connection, Device Mobility Info (DMI) for VPN attached phones should contain IP subnets distributed or owned by the VPN concentrators to ensure that "roaming" to a VPN location results in appropriate dynamic Device Mobility configuration changes. Be sure to associate the DMI with the same device pool that is used for any devices co-located with the VPN concentrators.
- If Device Mobility is enabled for users who access the enterprise network through Cisco Expressway mobile and remote access, Device Mobility Info (DMI) for Expressway attached devices should contain IP subnets used by the Expressway-C node(s) to ensure that "roaming" to an Expressway location results in appropriate dynamic Device Mobility configuration changes. Be sure to associate the DMI with the same device pool that is used for any devices co-located with the Expressway-C node.

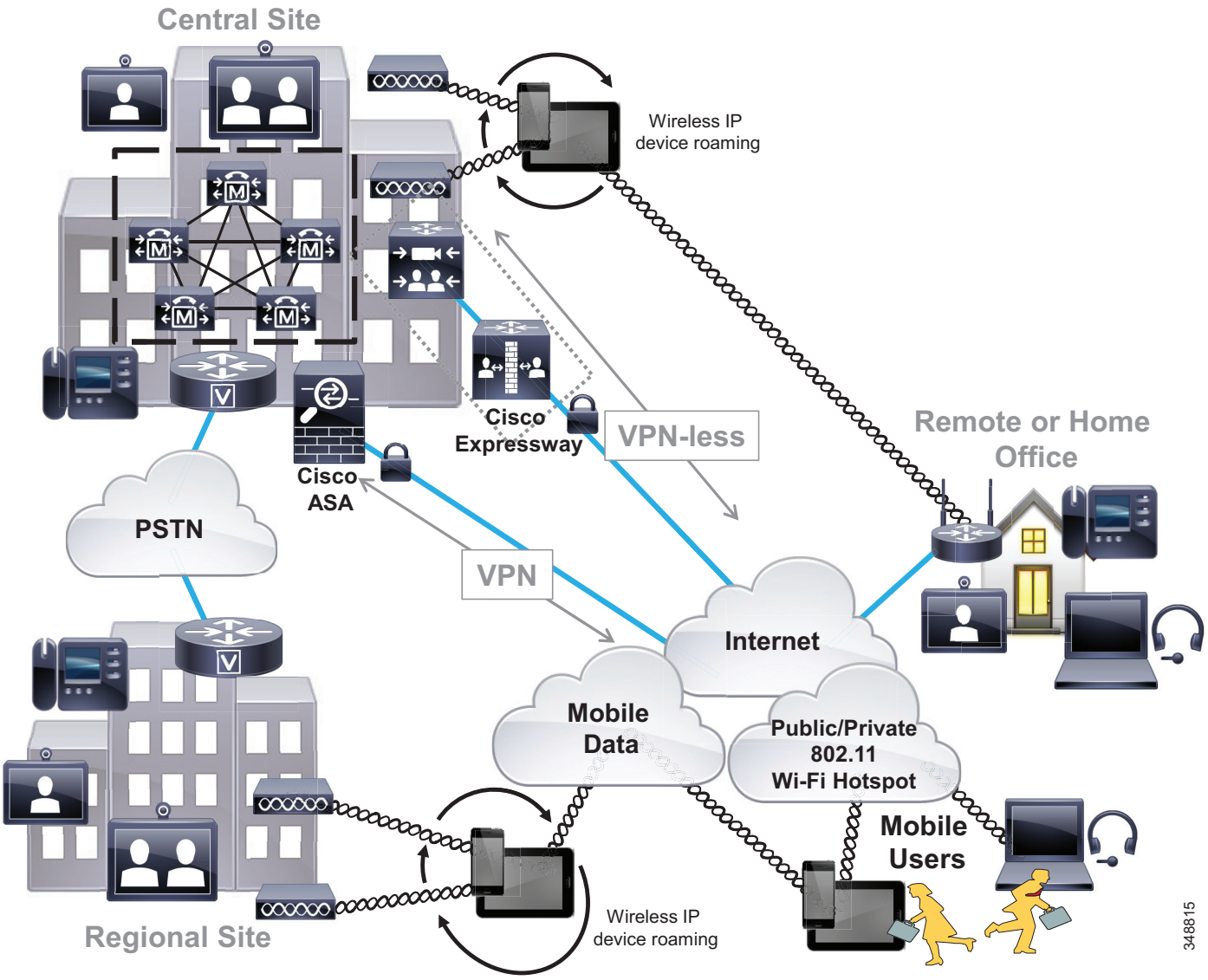
Remote Enterprise Mobility

Remote enterprise mobility refers to mobile users in locations remote from the enterprise but still attached to the enterprise network infrastructure through secure connections over the public Internet. Mobility here deals with the placement of endpoint devices in these remote locations and the movement of users, and in some cases their mobile devices, between the enterprise and these locations either frequently or on occasion.

Remote Enterprise Mobility Architecture

As illustrated in Figure 21-10, the remote enterprise mobility architecture is based on a remote physical location, typically an employee home office but also any remote location capable of secure connection back to the enterprise over the Internet. These remote sites typically consist of an IP network with connections for a user's computer, telephone, and other equipment or endpoints. In some cases this IP network may be behind an enterprise controlled and configured VPN router or edge security platform that provides a secure tunnel or connection between the remote location and the enterprise network. In other cases, the remote site IP network provides a connection to the Internet, and the user’s computer and other endpoint devices must use software-based client capabilities to create secure connections back to the enterprise network. Wireless connectivity may also be provided in the remote location to allow wireless attachment of the user's computer or endpoint. When wireless connectivity is provided at the remote location, wireless phones and mobile devices may be moved from the enterprise network to the home office, allowing users to leverage wireless enterprise devices or mobile phones within the remote location to make and receive calls.

Figure 21-10 Remote Enterprise Mobility Architecture



Types of Remote Enterprise Mobility

Remote enterprise mobility deployments focus predominately on supporting remote users as opposed to specifically supporting device mobility. Certainly users may regularly move with or without an endpoint device between the enterprise location or locations and remote sites; however, the predominate purpose of these deployments is to support remote connectivity for enterprise users, whether in a fixed location or in active motion. Remote site mobility involves two main types of secure remote connectivity, as shown in Figure 21-10:

- VPN secure remote connectivity
- VPN-less secure remote connectivity

VPN Secure Remote Connectivity

VPN secure remote connectivity enables a Layer 3 secure tunnel between the enterprise and the remote network or device. Using a VPN for secure remote enterprise connectivity in effect extends the boundary of the enterprise network to the VPN terminated location. VPN connections from VPN terminated devices or network locations provide network connectivity as though the device or network is located within the physical enterprise boundary. The Cisco Adaptive Security Appliance (ASA) head-end concentrator and Cisco AnyConnect clients enable VPN connectivity for both secure collaboration and other enterprise workflows. Router-based VPN connectivity and client-based VPN are the two common VPN deployment types. Both types support remote site secure connectivity and both can accommodate various endpoint devices, including both those remaining in a fixed location and those that can be moved between the remote site and the enterprise. Fixed location devices include wired video endpoints and IP phones as well as desktop computers. Dual-mode mobile phones, wireless IP phones, laptop computers, and tablets, are examples of endpoints that are mobile and are regularly moved between the remote site and the enterprise.

Router-Based Remote VPN Connectivity

Router-based VPN tunnels enable secure connectivity. As shown in Figure 21-10, in these types of scenarios the deployed remote site router (for example, the Cisco Virtual Office solution router) is responsible for setting up and securing a Layer 3 VPN tunnel back to the enterprise network. This in effect extends the enterprise network boundary to the remote site location. The advantage of this type of connectivity is that a wider range of devices and endpoints may be deployed in the remote site because these devices are not responsible for providing secure connectivity and therefore do not require special software or configuration. Instead, these devices simply connect to the remote site network and leverage the secure VPN IP path from the remote site router to the enterprise VPN head-end. The remote site router can also provide wireless network connectivity, as illustrated in Figure 21-10.

Client-Based Secure Remote Connectivity

Wireless and wired IP phones as well as software-based PC, smartphone, and tablet telephony clients can be connected over the Internet from remote network locations including home, mobile provider, and Wi-Fi hotspot networks, as shown in Figure 21-10. The VPN connection in the client-based VPN scenario is established through a software client running on the endpoint device. Thus the endpoint and software client are responsible for creating secure VPN connections back to the enterprise VPN head-end termination concentrator. This in effect extends the enterprise network boundary to the remote device. The advantage of this type of connectivity is that a wider range of network locations can be accommodated, including public networks where a router-based VPN connection is not practical. Connectivity across this diverse set of networks enables secure attachment while the client device is in motion. Depending on the endpoint device type, collaboration workflows such as voice and video calling might be the sole function leveraging the VPN connection. In the case of multipurpose devices such as PCs, smartphones, and tablets, full enterprise workflows are possible over the VPN connection.

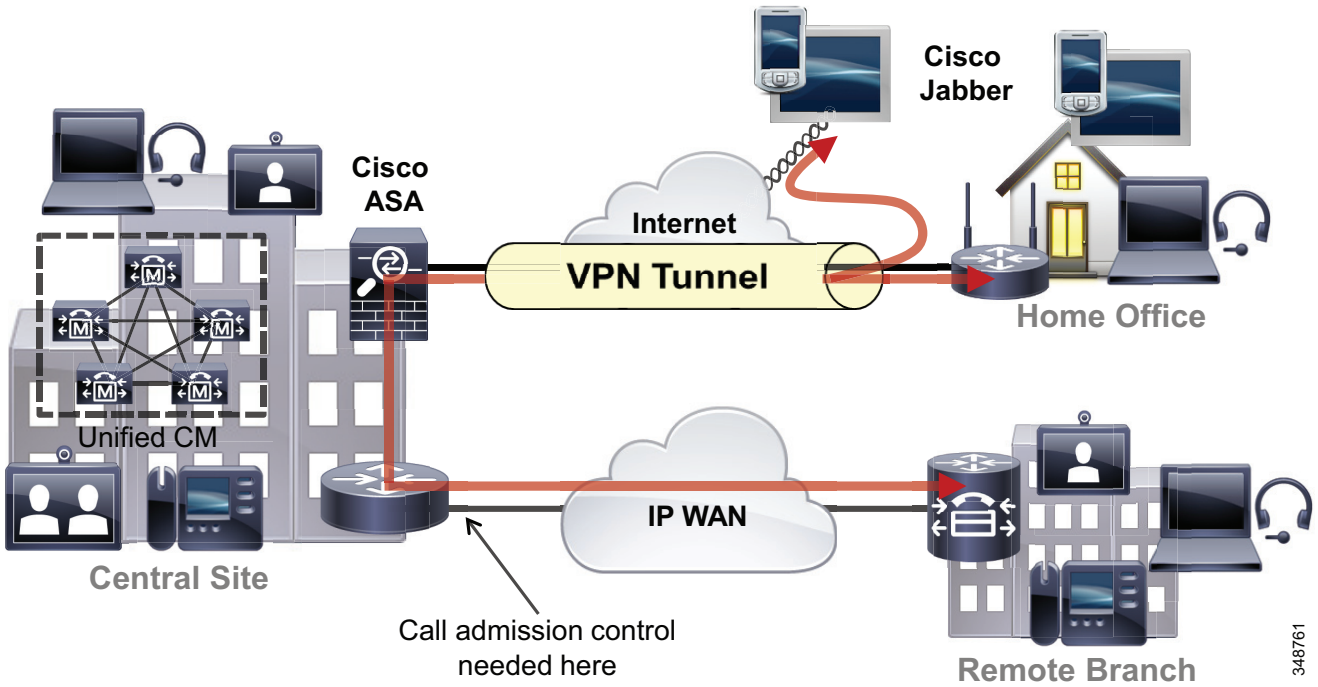
Examples of these types of devices include wired or wirelessly attached personal computers or wirelessly attached mobile client devices using a software-based VPN client such as the Cisco AnyConnect and wired Cisco Unified IP Phones such as the Cisco Unified IP Phone 7965, which uses a built-in VPN client.

Device Mobility and VPN Remote Enterprise Connectivity

Whether you are deploying client or router-based VPN remote connectivity, the Device Mobility feature may be used to ensure that call admission control and codec are correctly negotiated for endpoint devices and that the appropriate enterprise site PSTN gateway and media resources are utilized. Based on the IP address of the endpoint device as received over the VPN connection, Unified CM will dynamically determine the location of the device.

Figure 21-11 shows an example of client-based secure remote connectivity where a Cisco Jabber collaboration client is running on a remote site computer or mobile device. This software-based collaboration application is connected through a client-based VPN back to the enterprise and registered to Unified CM.

Figure 21-11 Client-Based VPN Connection for Remote Site Cisco Jabber



The following design guidelines pertain to enabling the Device Mobility feature for user devices at a remote site connected to the enterprise through a client or router-based VPN connection:

- Configure Device Mobility Info (DMI) with the IP subnets distributed or owned by the VPN concentrators.
- Associate the DMI with the same device pool that is used for devices co-located with the VPN concentrators. However, parameters such as calling privileges, network locale, and so forth, must be taken into consideration.
- Educate the remote site users to point to the geographically nearest enterprise VPN concentrator when making client-based or router-based VPN connections.

These guidelines ensure that call admission control is correctly applied on the enterprise WAN and over the connection to the remote site.

For information on deploying a VPN, refer to the various VPN design guides available under the *Security in WAN* subsection of the Design Zone for Security, available at:

https://www.cisco.com/c/en/us/solutions/enterprise/design-zone-security/landing_wan_security.html

VPN-Less Secure Remote Connectivity

VPN-less secure remote connectivity enables reverse proxy TLS secured connections between the enterprise and the remote attached device. This type of connectivity permits secure firewall traversal while minimizing the overhead required with a full Layer 3 VPN tunnel. Using a VPN-less reverse proxy secure connection extends the boundary of the enterprise network to the device or client application. The Cisco Collaboration Edge Architecture employs Cisco Expressway.

Cisco Expressway provides secure network traversal for specific endpoint or client application traffic flows as though this traffic is generated within the enterprise physical boundary. However, not all traffic flows are supported over this type of connectivity. The Cisco Collaboration Edge Architecture solution discussed here secures collaboration workflows including voice and video calling, IM and presence, visual voicemail, and corporate directory access. Full enterprise workflows including access to non-collaboration applications and services are not supported with these types of connections.

For more information on the Cisco Collaboration Edge Architecture, refer to the documentation available at

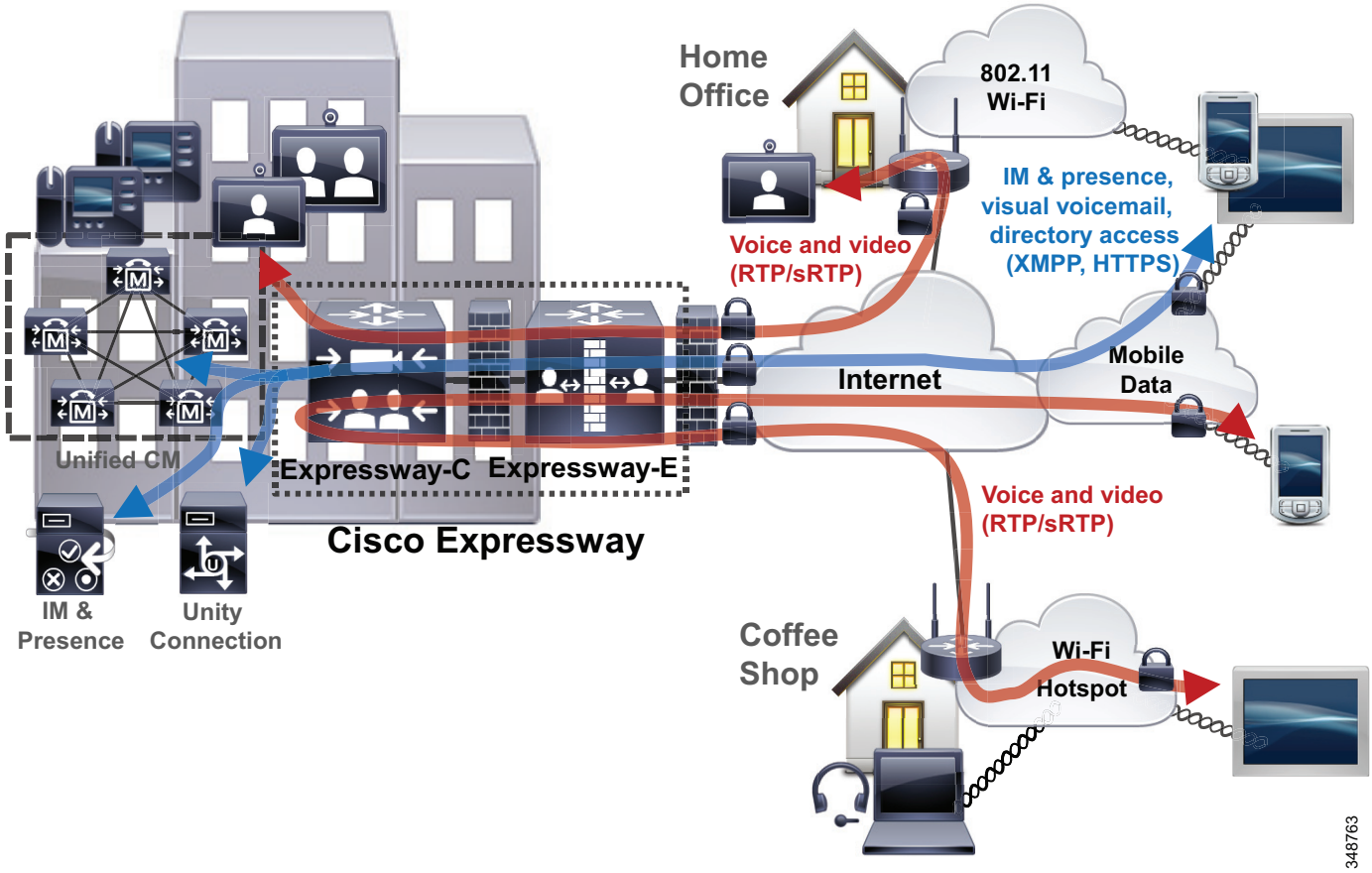
https://www.cisco.com/c/en/us/solutions/collaboration/collaboration-edge-architecture/index.html

Cisco Expressway

The mobile and remote access feature of the Cisco Expressway solution provides secure reverse proxy firewall traversal connectivity, which enables remote users and their devices to access and consume enterprise collaboration applications and services.

As shown in Figure 21-12, the Cisco Expressway solution encompasses two main components: the Expressway-E node and the Expressway-C node. These two components work in combination with Unified CM to enable secure mobile and remote access. The Expressway-E node provides the secure edge interface to mobile and remote devices. This node normally resides in the DMZ area of the enterprise network. The Expressway-C node provides proxy registration to Unified CM for remote secure endpoint registration. The Expressway-C node, residing in the internal enterprise network, creates an outbound secure TLS connection with the Expressway-E node, and this connection is leveraged for secure media traversal.

Figure 21-12 Secure Remote Collaboration with Cisco Expressway Mobile and Remote Access



Once registered to Unified CM, the remote device is able to make and receive voice and video calls over IP using SIP signaling and RTP media. The secure Cisco Expressway mobile and remote connection not only enables device registration and voice and video calling, but it also enables additional collaboration workflows including IM and presence, visual voicemail, and corporate directory access. The full collaboration feature set is available from the enterprise without requiring a VPN tunnel. Voice and video media as well as signaling and other collaboration traffic traverse the enterprise network at the Expressway-C node. As shown in Figure 21-12, calls between two remote devices outside the enterprise will be hairpinned at the Expressway-C node within the enterprise.

Unlike with VPN secure connections where all traffic from the secured endpoint traverses the VPN tunnel back to the enterprise, Cisco Expressway mobile and remote access enables secure connectivity to the enterprise for collaboration traffic only. Non-collaboration workflows and traffic do not traverse the secure Cisco Expressway connection. Instead, all other traffic is sent directly to the local network or the Internet and does not traverse the enterprise network.

The Cisco Expressway mobile and remote access functionality supports both Cisco hardware endpoints and Cisco Jabber software-based client endpoints. Supported Cisco hardware endpoints include Cisco TelePresence EX, MX, and SX Series video endpoints and Cisco DX, 7800, and 8800 Series desk phones. Cisco Jabber desktop and mobile clients also support Cisco Expressway mobile and remote

access. In particular, Cisco Jabber mobile clients support Cisco Expressway mobile and remote access connectivity while in motion, thus enabling secure real-time collaboration regardless of the mobile user's location or network connectivity type.

Just as when relying on VPN for remote secure connectivity, Device Mobility configuration with Expressway mobile and remote access is critical for ensuring Unified CM is able to track endpoint locations for the purposes of monitoring call volume over low-speed links, negotiating appropriate codecs, and routing calls using local gateway resources. When configuring Device Mobility in environments with Expressway mobile and remote access, remember to:

- Configure Device Mobility Info (DMI) with the IP subnet(s) used by the Expressway-C nodes.
- Associate the DMI with the same device pool that is used for any devices co-located with the Expressway-C node.

Cisco Expressway mobile and remote access functionality supports a maximum of 10,000 remote endpoint registrations to Unified CM per Expressway-C and Expressway-E cluster pair. In addition, Expressway cluster pairs support a maximum of 2,000 simultaneous video calls or 4,000 simultaneous voice-only calls. For more information about Cisco Expressway capacity, including per-Expressway node capacities, see the section on Cisco Expressway, page 25-37.

Deploy multiple Expressway clusters for increased scale or for designs spanning multiple geographic locations. In the case of multi-site deployments, Expressway clusters should be distributed across geographic regions to provide remote enterprise connectivity to users and their devices regardless of location. In order to effectively distribute Expressway mobile and remote access connections so that devices connect to the nearest Expressway service node or cluster, GeoDNS services are recommended. With GeoDNS service, mobile devices are usually directed to the nearest Expressway service point based on location as determined by the source IP address of the DNS query for Expressway DNS service records or based on the shortest mean latency between the location of the device and available Expressway service nodes.

For more information about the Cisco Expressway solution, refer to the data sheet and documentation available at

https://www.cisco.com/c/en/us/products/unified-communications/expressway-series/index.html

Remote Enterprise Mobility High Availability

For remote site mobility environments, it is imperative that enterprise VPN or VPN-less security services are configured and deployed in a redundant manner within the enterprise. This ensures that VPN and reverse proxy firewall traversal secure connections are highly available. If a VPN concentrator or Cisco Expressway node within the enterprise or at the enterprise edge fails, a new secure connection can be set up by the client or endpoint with another VPN or VPN-less remote edge node. Device registration, voice and video services, IM and presence, and other collaboration services are highly available based on the Unified CM cluster or other application server node redundancy. This level of collaboration service redundancy applies on-premises as well as when endpoints and clients are connected to the enterprise through a VPN.

Collaboration application and service redundancy is limited when endpoints and clients connect using Cisco Expressway mobile and remote access. In the case of the Cisco Expressway solution, high availability for mobile and remote access is achieved by deploying clusters of each node type. In a deployment with a cluster of Unified CM nodes, a cluster of Expressway-E nodes, and a cluster of Expressway-C nodes, backup nodes are able to provide mobile and remote access and device registration in scenarios where one or more primary nodes fail.

Capacity Planning for Remote Enterprise Mobility

The most critical scalability consideration for remote enterprise mobility environments is the enterprise head-end session terminator. Administrators must deploy sufficient VPN session and VPN-less connectivity capacity to accommodate all remote secure attachment requirements. Whether client or router-based VPN or VPN-less remote edge secure connections through Cisco Expressway, in all cases sufficient platform or node capacity must be provided to handle the device registration load as well as the various collaboration workflows available over the secure connection. Failure to provide appropriate capacity will prevent some remote sites and devices from connecting to the enterprise, thus eliminating access to even basic telephony services. Furthermore, just as with campus and multisite enterprise mobility deployments, it is important to provide sufficient device registration capacity within the enterprise to handle all remote user devices.

For more information on Cisco call control and gateway edge capacity, including platform-specific endpoint configuration and registration capacities, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Design Considerations for Remote Enterprise Mobility

Consider the following design recommendations when enabling remote site connectivity for mobile users:

- When using Device Mobility, remember to configure Device Mobility Info (DMI) with the IP subnets distributed or owned by the VPN concentrators, or in the case of Expressway, with the subnet(s) used by the Expressway-C nodes. Assign the DMI to the same device pool that is configured for devices deployed in the same location as the VPN concentrators or Expressway-C nodes.
- Educate remote site users to select the nearest VPN concentrator for VPN connection.
- Ensure appropriate VPN session capacity is available in order to provide connectivity to all remote site locations and devices using VPN.
- Ensure appropriate reverse proxy firewall traversal session capacity is available in order to provide VPN-less secure connectivity to all remote devices. Ensure that sufficient Expressway-E and Expressway-C nodes and session capacity are available. In all cases, sufficient Unified CM registration capacity is required.

Cloud and Hybrid Services Mobility

Cloud and hybrid services mobility refers to mobile users utilizing collaboration applications and services delivered from the Cisco Collaboration cloud. This type of mobility includes both pure cloud deployments leveraging only cloud collaboration services and hybrid deployments leveraging both cloud and enterprise on-premises collaboration applications and services.

Mobile devices and clients connect over the Internet to the Cisco Collaboration Cloud and other cloud collaboration applications and services. Clients and devices can be located either on-premises or remotely from the enterprise. With access to the Internet, devices (whether in motion or at rest) can consume these services connected through the enterprise network or through a public or private network.

Enterprises choose to enable collaboration services from the cloud and in some cases integrate these services with the enterprise collaboration infrastructure for a variety of reasons. The main reasons enterprises increasingly look to the cloud for delivering software services and applications are:

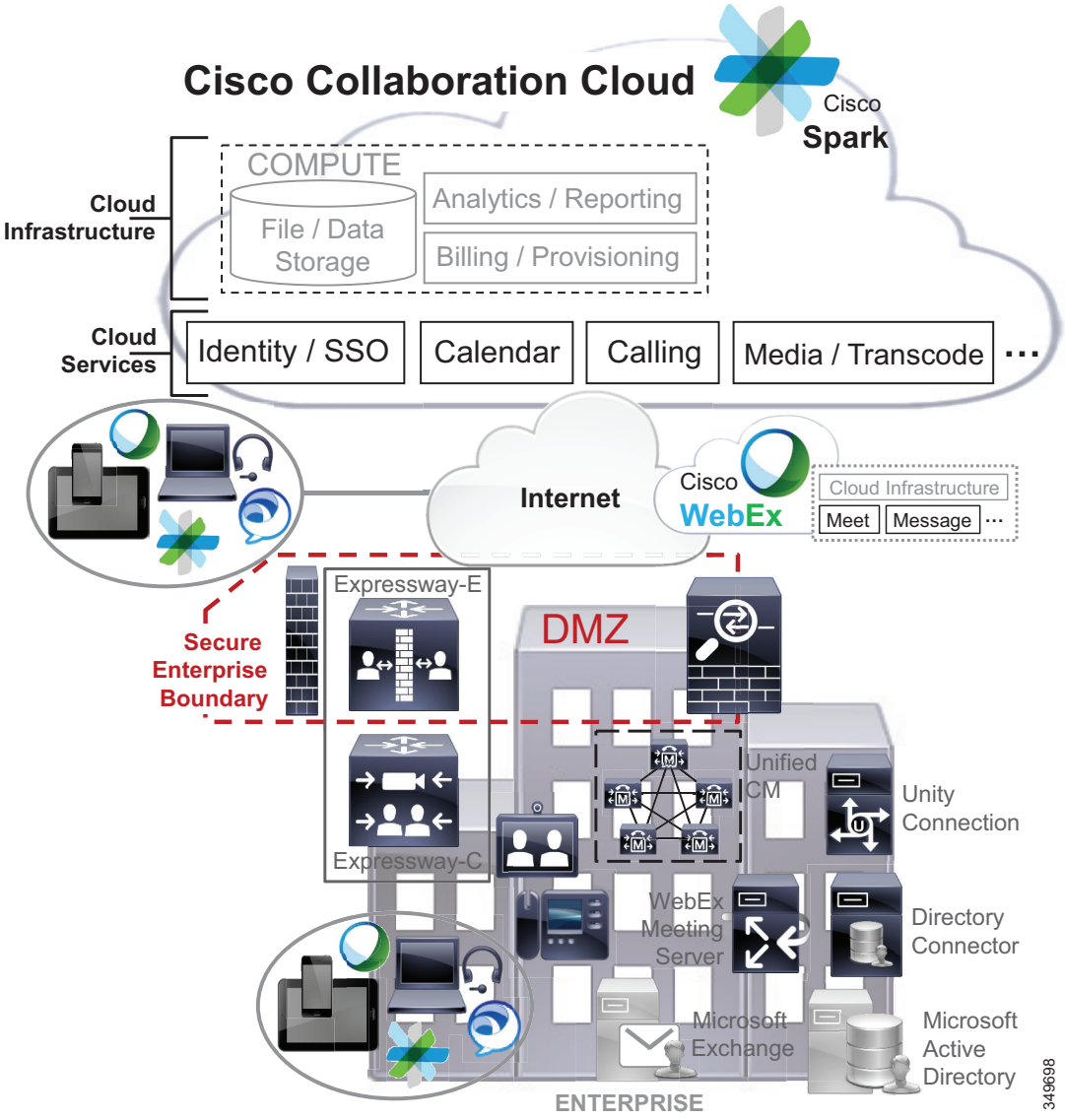
- Continuous and automatic delivery of cloud service updates to provide rapid deployment of new features and fixes to resolve reported issues
- Elasticity of compute resources, enabling on-demand user capacity and service performance
- Centralized on-line administration and management of cloud application and service features and functions
- Highly available cloud architecture, providing geographic coverage and service resiliency
- Infrastructure capital expenditure and management off-loaded to the cloud vendor. The vendor manages and secures the infrastructure, including compute, storage, power, network, and foundational services and applications.

Cloud and Hybrid Service Mobility Architecture

As illustrated in Figure 21-13, the cloud and hybrid service mobility architecture is based on the Cisco Collaboration Cloud and Cisco WebEx Collaboration Cloud services connected to the Internet. The Collaboration Cloud and WebEx Collaboration Cloud services are enabled on an underlying secure, resilient cloud compute infrastructure. Cloud collaboration services delivered with this architecture include Cisco Spark message, meet, and call, and WebEx meetings and messaging. In addition to pure cloud deployments of these services, they may also be deployed in conjunction with enterprise on-premises services. For example, an enterprise may enable WebEx Meeting Center meetings and WebEx Messenger IM and presence (services from the cloud) in tandem with Unified CM voice and video calling and Unity Connection voice messaging (services delivered on-premises). Cisco Spark message, meet, and calling may be augmented with cloud hybrid service enterprise integrations including enterprise identity, single sign-on (SSO), calendaring, and calling.

Cloud service enterprise integrations generally rely on secure connections between the cloud and the enterprise to transmit service-related traffic to and from the enterprise. This traffic must traverse the secure enterprise boundary DMZ, as shown in Figure 21-13.

Figure 21-13 Cloud and Hybrid Services Mobility Architecture



Cisco desktop, web browser, and mobile device collaboration applications and clients – including Cisco Jabber, Cisco Spark, and Cisco WebEx – consume services from the Cisco Collaboration and WebEx Collaboration clouds, whether connected remotely through the Internet when outside the enterprise or connected from within the enterprise.

For more information about Cisco clients capable of leveraging cloud-based services, see Cisco Mobile Clients and Devices, page 21-76.

Types of Cloud Hybrid Service Integrations

There are two primary types of cloud hybrid collaboration service integrations:

- Cisco WebEx Collaboration Cloud Hybrid Integrations, page 21-36
- Cisco Spark Hybrid Services, page 21-36

Cisco WebEx Collaboration Cloud Hybrid Integrations

While Cisco WebEx collaboration cloud capabilities are available as standalone services, they can also augment existing enterprise on-premises collaboration services through hybrid integrations to enable:

- Instant messaging (IM) and presence with the Cisco WebEx Messenger service
- Voice and video conferencing with desktop sharing with Cisco WebEx Meetings services.

Cisco WebEx hybrid integrations are not covered in this chapter.

For information on the Cisco WebEx Collaboration Cloud and hybrid enterprise collaboration integrations, see the section on Cisco WebEx Software as a Service, page 11-26.

For information on Cisco WebEx Messenger and hybrid enterprise integrations, see the section on Cisco WebEx Messenger, page 20-64.

Cisco Spark Hybrid Services

The Cisco Spark hybrid collaboration service integrations enabled for the Cisco Collaboration Cloud include:

- Cisco Spark Identity Service, page 21-36
- Cisco Spark Calendar Service, page 21-38
- Cisco Spark Call Service, page 21-41

For general information about Cisco Spark Hybrid Services, refer to the introductory information at https://collaborationhelp.cisco.com/article/en-us/DOC-6433.

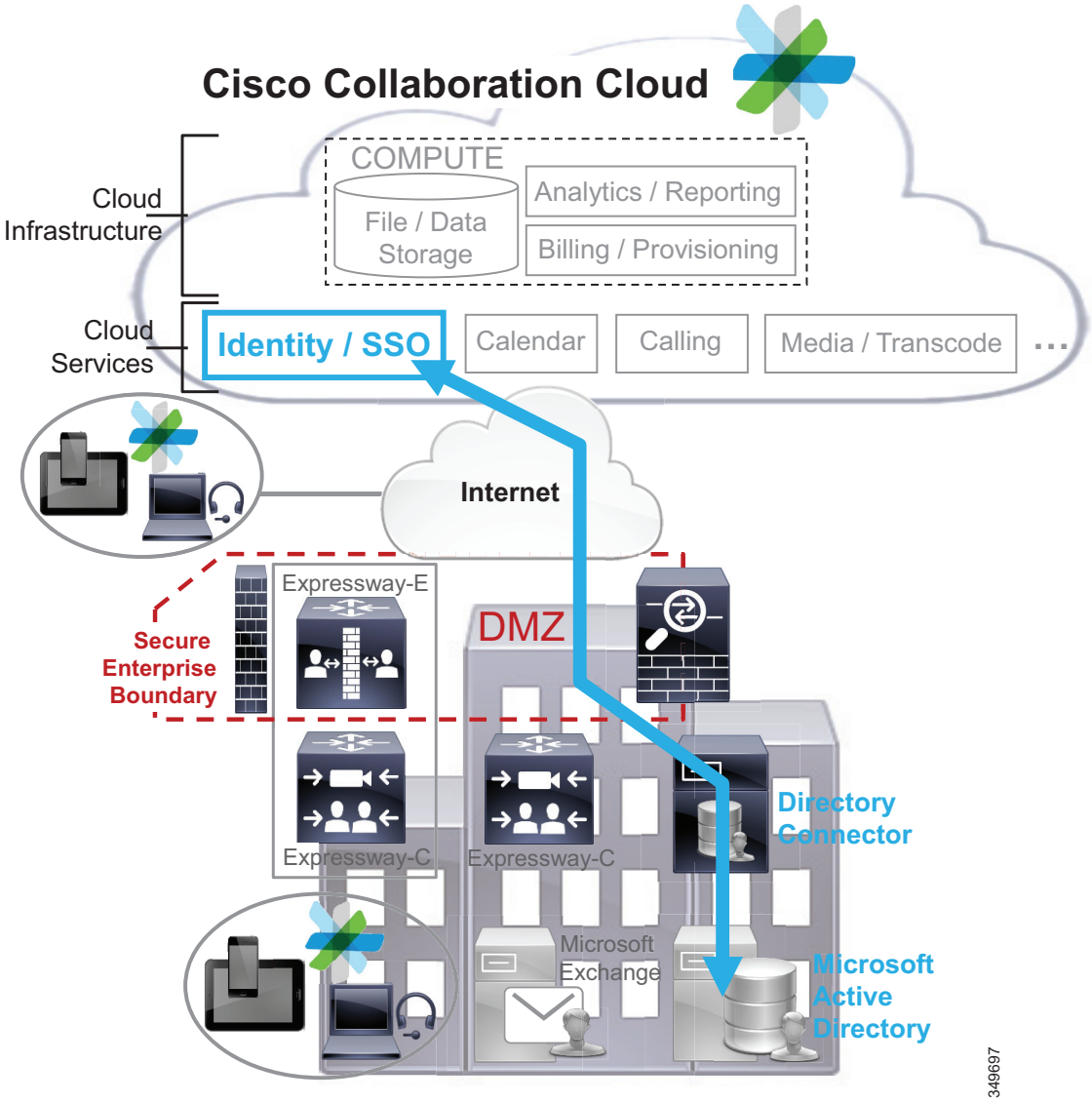
Cisco Spark Identity Service

Cisco Spark Hybrid Services provide a mechanism for integrating on-premises enterprise Microsoft Active Directory with the Cisco Collaboration Cloud Common Identity Services (CIS). By syncing enterprise directory information with CIS in the cloud, organizations can enable rapid configuration and provisioning of enterprise users for Cisco Cloud. Note that cloud identity services also include single sign-on (SSO) capabilities, should the enterprise wish to implement or integrate SSO for Cisco Spark Hybrid Services.

As shown in Figure 21-14, the on-premises Cisco Directory Connector communicates and synchronizes over the enterprise network with Microsoft Active Directory. In turn, the Directory Connector pushes directory data and communicates over the Internet through the secure enterprise boundary and corporate firewall with the cloud identity (CIS) and SSO service. This connection is initiated from inside the enterprise to the cloud and does not require ports to be opened on the corporate firewall. This is similar to an HTTPS web client that initiates an outbound connection to a web server on the Internet and receives a response on that same connection.

HTTPS is used for communication between CIS in the cloud and the on-premises Cisco Directory Connector. Microsoft Active Directory APIs are used for synchronization between the Cisco Directory Connector and Microsoft Active Directory.

Figure 21-14 Cisco Spark Hybrid Services: Cloud Identity Service and Enterprise Directory Integration



The connection between CIS and Directory Connector used to synchronize users is set up automatically during installation of the Directory Connector software. The Directory Connector software is downloaded from the Cisco Spark Control Hub. The connection between the Directory Connector and Microsoft AD used to synchronize user information is controlled by configuration on the Directory Connector. Configure object types, LDAP field mappings, and base DN(s) using the Directory Connector administrative graphical user interface to control which user accounts and what account information are synchronized.

The Cisco Directory Connector software installs and runs on a server or virtual machine with the Microsoft Windows Server operating system. The following requirements and recommendations apply to the Cisco Directory Connector deployment:

- The Microsoft Windows server or virtual machine must be a member of the enterprise Microsoft Active Directory domain.
- The Directory Connector software must be installed on the Windows server or virtual machine using an account with domain administration privileges.
- The email address attribute in Active Directory must be populated for all user accounts to be synchronized with Cisco CIS. User accounts in Active Directory without an email address will not be synced with CIS.
- We recommend that you install the Directory Connector on a server or virtual machine separate from the Active Directory Domain Service (AD DS) and Active Directory Lightweight Directory Services (AD LDS).

For more information about the Cisco Directory Connector, including deployment requirements, installation, and configuration, refer to latest version of the *Deployment Guide for Cisco Directory Connector*, available at

https://www.cisco.com/c/en/us/support/unified-communications/spark/products-installation-guides-list.html

Once enterprise users are synchronized between the on-premises Microsoft Active Directory and the Cisco Collaboration Cloud CIS, the organization administrator is easily able to manage user accounts using the Cisco Spark Control Hub. From this hub the administrator assigns user roles, manages user capabilities, and entitles or activates users for specific cloud services, including Cisco Spark Hybrid Services.

Cisco Spark Calendar Service

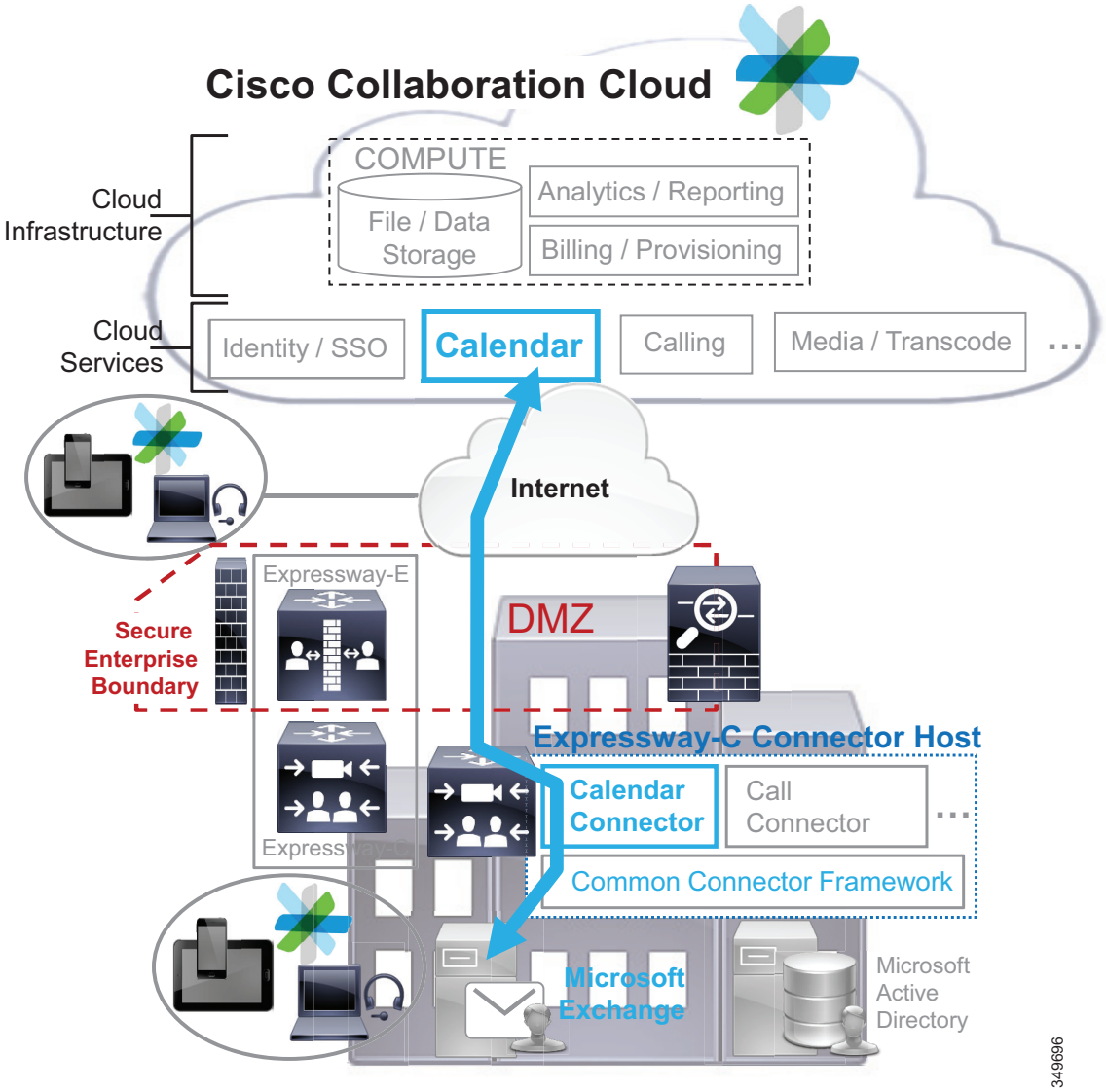
Cisco Spark Hybrid Services provide a mechanism for integrating on-premises enterprise Microsoft Exchange calendaring capabilities with the Cisco Collaboration Cloud calendar service. With enterprise calendar service integration to the Cisco Collaboration Cloud, organizations can automatically incorporate the rich collaboration capabilities of Cisco Spark and Cisco WebEx into their Outlook meeting invitations by simply including @spark and/or @webex in the meeting invitation location field.

The Calendar Connector is responsible for brokering the integration and communication between the cloud calendar service and the enterprise Exchange environment. As shown in Figure 21-15, the on-premises Cisco Expressway-C Connector Host Calendar Connector, relying on the underlying Common Connector Framework, communicates with Microsoft Exchange over the enterprise network. In turn, the Calendar Connector pushes calendar data and communicates over the Internet through the secure enterprise boundary and corporate firewall to the cloud calendar service. This connection is initiated from inside the enterprise to the cloud and does not require ports to be opened on the corporate firewall. This is similar to an HTTPS web client that initiates an outbound connection to a web server on the Internet and receives a response on that same connection.

HTTPS is used for communication between the calendar service in the cloud and the on-premises Calendar Connector. Microsoft Exchange Web Services (EWS) are used for communication between the Expressway-C Calendar Connector component and the Microsoft Exchange environment.

The Calendar Connector communicates with the Exchange environment to monitor notifications and retrieve information from users' calendars and to add Cisco Spark room and WebEx meeting information to meeting invitations.

Figure 21-15 Cisco Spark Hybrid Services: Enterprise Calendar Integration



The connection between the Expressway-C Connector Host and the Cisco Collaboration Cloud, and the connection between the cloud calendar service and the Calendar Connector, are established automatically during configuration of hybrid services connectors on the Expressway-C. The Calendar Connector software is automatically downloaded and installed on Expressway-C from the Cisco Collaboration Cloud following successful Expressway-C registration (or if registration has already occurred, when the calendar connector service is activated from the Cisco Spark Control Hub).

During configuration of the Calendar Connector through the Expressway-C graphical user interface, the Microsoft Exchange connection information is provided by the administrator (or alternatively may be retrieved from the enterprise Active Directory). The administrator also specifies the organization's WebEx Meeting Center and Collaboration Meeting Room site information so that WebEx meeting room details are added to meeting invitations when @webex is specified in the invitation location field.

Cisco Spark Hybrid Services that rely on Expressway-C and the Common Connector Framework require a secure connection between the Cisco Collaboration Cloud and the on-premises Expressway-C. In order for the Expressway-C Calendar Connector to operate, the CA-signed certificates offered by the Cisco Collaboration Cloud for connector management and calendar service are verified against the Expressway-C certificate trust list. This provides a secure connection between Expressway-C and the Collaboration Cloud. Expressway-C verifies cloud certificates prior to downloading the Calendar Connector software and starting the Calendar Connector service. The Calendar Connector service will not start if the cloud certificate CA is not in the trust list. The cloud will automatically append the required cloud public CA certificates to the Expressway-C trust list during initial configuration. Alternatively, organizations may choose to manage cloud certificates manually, in which case the Expressway administrator must append cloud CA certificates to the Expressway trust list for proper operation. Secure connectivity is optionally extended to the connection between the Expressway-C Calendar Connector and the enterprise Exchange server by exchanging CA certificates and appending to the trust list of the respective servers.

For proper integration and communication between the Calendar Connector and Microsoft Exchange, an impersonation account must be used. This account is used by Calendar Connector on behalf of users to query their individual calendars for meeting information. The Calendar Connector does not use this account to access user email or contact lists, and the Cisco Collaboration Cloud is not able to access or retrieve the Exchange environment impersonation account credentials from the connector. Further, the Collaboration Cloud has no access, directly or through the Calendar Connector, to the enterprise Exchange environment.

The following requirements and recommendations apply to the Calendar Connector deployments:

- Because hybrid services users are authenticated against the Collaboration Cloud Common Identity Service (CIS), Cisco Directory Connector and integration to the enterprise Active Directory are recommended.
- Cisco Expressway X8.7.1 or later version is required for Cisco Spark Hybrid Services.
- The number of users entitled for calendar service, the size of individual user Exchange calendars, and the rate at which @spark and @webex are used, will determine the amount of increased load on the Exchange server when enabling this service. Create and apply a throttling policy on the Exchange impersonation account to reduce the impact of the Calendar Connector and calendar services on the enterprise Exchange environment.

For more information about the Calendar Connector, including deployment requirements, installation, and configuration, refer to the latest version of the *Deployment Guide for Cisco Spark Hybrid Calendar Service*, available at

https://www.cisco.com/c/en/us/support/unified-communications/spark/products-installation-guides-list.html

Once Calendar Connector is running and users are activated, enabled users can incorporate Cisco Spark collaboration and add WebEx meeting information to Outlook calendar invitations by including the following:

- @spark

  When @spark is added to the location field of an Outlook calendar invitation, Calendar Connector and the cloud calendar service create a new Cisco Spark collaboration room with a name that matches the invitation subject. All users in the calendar invitation are added to the Cisco Spark room. This facilitates collaboration and allows the meeting organizer and attendees to communicate and share material prior to, during, and even after the meeting. If a calendar invitation includes a distribution list, users on the distribution list will not be added to the Cisco Spark room automatically; however, they will receive the meeting invitation.

- @webex — When specified, it adds WebEx meeting invitation information into the Cisco Spark room.

  When @webex (or @webex:<*site*> for organizations with multiple WebEx sites) is added to the location field of an Outlook calendar invitation, Calendar Connector automatically populates the invitation with the user's WebEx collaboration meeting room information. Calendar Connector will not add WebEx meeting information if any WebEx meeting join links (added manually or by WebEx Productivity tools) are already present in the calendar invitation.

When @webex is used in conjunction with @spark, WebEx meeting information is added to the Cisco Spark room as well as the calendar meeting invitation.

Cisco Spark Call Service

Cisco Spark Hybrid Services enable integration of the Cisco Collaboration Cloud calling service with on-premises enterprise call control. With enterprise call service integration to the cloud, an organization can enable desktop sharing and voice and video calling between existing on-premises phones and collaboration clients and Cisco Spark clients.

As shown in Figure 21-16, there are three enterprise components required for Cisco Spark hybrid call service:

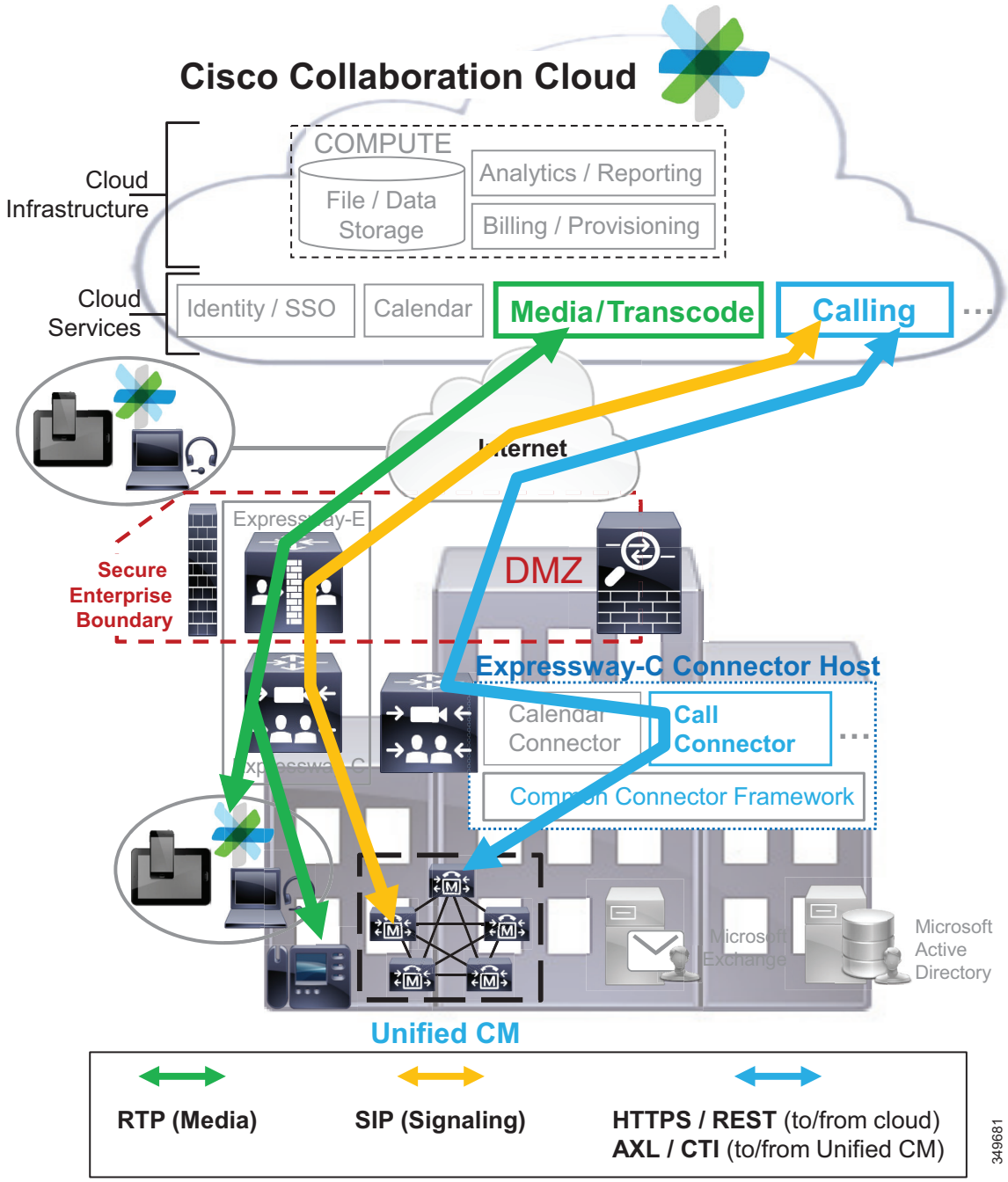
- Cisco Call Connector — This software runs on the Cisco Expressway-C Connector Host and brokers the integration and communication between the Cisco Collaboration Cloud calling service and the enterprise Unified CM deployment.
- Cisco Unified CM — This is the enterprise call control, and it provides voice and video calling services and PSTN connectivity for enterprise endpoints and clients and enterprise-connected cloud clients. Enterprise call control may also be provided by Cisco Business Edition 6000 or Cisco Hosted Collaboration System (HCS).
- Cisco Expressway-E and Expressway-C — These server pairs provide secure enterprise edge firewall traversal for call media and signaling. Existing server pairs used for Expressway mobile and remote access or business-to-business (B2B) may be leveraged if sufficient call capacity is available.

The Call Connector residing on the Cisco Expressway-C Connector Host relies on the underlying Common Connector Framework to communicate with Unified CM over the enterprise network. As with other enterprise cloud connectors, the Call Connector communicates over the Internet through the secure enterprise boundary and corporate firewall to the cloud. This connection is initiated from inside the enterprise to the cloud and does not require ports to be opened on the corporate firewall. As mentioned previously, this is similar to an HTTPS web client initiating an outbound connection to a web server on the Internet.

The Call Connector communicates with the Cisco Collaboration Cloud calling service using REST-based HTTPS. It communicates with Unified CM using Administrative XML Layer (AXL) to retrieve a user's enterprise device information and using Computer Telephony Integration (CTI) to monitor the user's enterprise line.

Just as with the Calendar Connector, the connections between the Expressway-C Connector Host, the Call Connector, the Cisco Collaboration Cloud, and the cloud calling service, are established automatically during configuration of hybrid services connectors on the Expressway-C. The Call Connector software is automatically downloaded and installed on Connector Host from the Cisco Collaboration Cloud following successful Expressway-C Connector Host registration (or if registration has already occurred, when the Call Connector service is activated from both the Cisco Spark Control Hub and the Expressway-C Connector Host).

Figure 21-16 Cisco Spark Hybrid Services: Enterprise Calling Integration



Cisco Spark Hybrid Services calling enables two features:

- Call Service Aware

  This feature provides one-click-to-share capabilities for calls between two Cisco Spark-enabled users on their Unified CM registered endpoints. When two users are in a one-on-one call using their enterprise line, the cloud calling service is aware of the active call based on information received from the hybrid service Call Connector, and it automatically brings the Cisco Spark room between the two users to the top of the list (or creates a one-on-one room if one has not been created previously) and enables a desktop share button within the room on both users' Cisco Spark desktop (or web) client. Either user can click the button to share their desktop. With Call Service Aware, call media and signaling for the one-on-one call is handled exclusively by Unified CM and the two enterprise devices, while the Cisco Collaboration Cloud facilitates the Cisco Spark collaboration room and desktop share. Besides enabling desktop share, Call Service Aware also provides a unified call history list for Cisco Spark clients.

- Call Service Connect

  This feature enables Cisco Spark users to make and receive calls using the on-premises enterprise call control (Cisco Unified CM). When this feature is configured, an incoming call to a user's enterprise number is not only extended to the user's Unified CM registered phones and clients, but it is also extended to the Cisco Collaboration Cloud and routed to the user's Cisco Spark client(s), thus allowing the user to answer the call using their most readily available device whether that is, for example, an enterprise registered desk phone or the Cisco Spark client running on the user's mobile phone. Likewise, incoming Cisco Spark originated calls to the user are not only extended to the user's Cisco Spark client but are also extended by the Cisco Collaboration Cloud to the enterprise Unified CM and ring in on the user's Unified CM registered endpoints.

  In cases where a user makes a call by entering a number or URI within the Cisco Spark client calls tab, the call is routed using the enterprise Unified CM and the enterprise PSTN connection if needed. Call Service Connect cannot be enabled for the user without the Call Service Aware feature being enabled as well.

  The Call Service Connect feature requires each user to have a Cisco Spark Remote Device (Spark RD) configured within Unified CM. This device associates the Cisco Collaboration Cloud user with an enterprise DN and a Cisco Spark calling SIP URI configured as a remote destination. This device association and the configured remote destination facilitate call forking to both Unified CM and the Collaboration Cloud, depending on where the call originates. The Cisco Collaboration Cloud uses SIP contact headers and call routing logic to prevent call forking loops between the cloud and enterprise call control.

  Note Early deployments of Cisco Spark Call Service Connect used the CTI Remote Device; however, the proper Unified CM device type for current deployments is the Cisco Spark Remote Device (Spark RD).

  The Call Connector registers the Spark RD with Unified CM. This registration is active as long as the Call Connector is connected to the Cisco Collaboration Cloud

  With Call Service Connect enabled, RTP call media and SIP call signaling are routed to and from the Cisco Collaboration Cloud using Expressway-E and Expressway-C server pairs as illustrated in Figure 21-16. Call media traverses the Expressway-E and Expressway-C servers between the enterprise attached endpoint (or gateway) and the Cisco Collaboration Cloud media and transcoding service. SIP signaling between the cloud calling service and Unified CM also traverses the Expressway-E and Expressway-C servers. RTP media and SIP signaling for Call Service Connect can traverse existing mobile and remote access or B2B Expressway-E and Expressway-C servers, or a dedicated set of Hybrid Services Expressway-E and Expressway-C servers may be deployed.

Just as with Cisco Spark Calendar Service, Cisco Spark Calling Service also relies on a secure connection between the Expressway-C Connector Host and the Common Connector Framework. And just as with Calendar Connector, in order for Call Connector to operate, the CA-signed certificates offered by the Cisco Collaboration Cloud for connector management and call service are verified against the Expressway-C Connector Host certificate trust list. The Expressway-C Connector Host verifies cloud certificates prior to downloading the Call Connector software and starting the connector service. The Call Connector service will not start if the cloud certificate CA is not in the trust list. The cloud automatically appends the required cloud public CA certificates to the Expressway-C Connector Host trust list during initial configuration. Alternatively, cloud certificates can be managed manually, requiring the administrator to append cloud certificates to the Expressway-C Connector Host certificate trust list.

The following requirements and recommendations apply to the Call Connector deployments:

- Because hybrid services users are authenticated against the Collaboration Cloud Common Identity Service (CIS), Cisco Directory Connector and integration to the enterprise Active Directory is required.
- Cisco Expressway X8.7.1 or later version is required for Cisco Spark Hybrid Services.
- Call Service Aware is a prerequisite for the Call Service Connect feature.
- The AXL Web Service and CTIManager services required for Cisco Spark Call Service should be enabled on at least two Unified CM nodes to provide high availability.

For more information about the Call Connector, including deployment requirements, installation, and configuration, refer to the Call Service Aware setup information in the latest version of the *Deployment Guide for Cisco Spark Hybrid Call Services*, available at

https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-installation-and-configuration-guides-list.html

Cloud and Hybrid Services Mobility High Availability

Like other enterprise mobility features and solutions, cloud and hybrid services should be configured and deployed in a redundant fashion to provide high availability of cloud services. By their nature, cloud infrastructures and platforms are resilient. As with most managed cloud infrastructures, the Cisco Collaboration Cloud and WebEx Cloud rely on sophisticated RAID storage arrays and power grids, continuous data backup, and on-demand computing with data center distribution and migration capabilities to ensure highly available cloud services.

In the case of hybrid service deployments, besides cloud resiliency, on-premises infrastructure redundancy must also be considered. It is critical to deploy on-premises enterprise network infrastructure components, including the enterprise network and secure enterprise boundary, in a highly available fashion. Collaboration components, including WebEx Meeting Server, Expressway-C Connector Host, and enterprise applications such as Microsoft Exchange and Active Directory, should be deployed in a redundant fashion.

Traditional Microsoft Exchange and Active Directory high availability deployment methods are likely in place, assuming these applications are critical for enterprise operation. If not, consider implementing high availability for these applications. On-premises Microsoft application high availability also applies to hybrid service integrations.

Capacity Planning for Cloud and Hybrid Services Mobility

Deploying cloud and hybrid services successfully requires ample capacity to accommodate all users that will utilize the cloud services. While cloud capacity is on-demand and virtually limitless given the elastic nature of cloud computing and storage, the cost of entitlement needs to be considered.

Hybrid integrations introduce additional scalability considerations given the enterprise on-premises infrastructure. In the case of Microsoft applications (Exchange and Active Directory), follow Microsoft guidance related to capacity and ensure that appropriate capacity is provided for the additional overhead of hybrid services beyond the existing on-premises utilization. In particular it is important to implement a throttling policy on the Exchange server to prevent over-subscription of server resources.

The Expressway-C node (large OVA or large appliance) supports a maximum of 5,000 cloud hybrid service users.

Also, in the case of Directory Connector, enterprises planning to synchronize large numbers of users with the Collaboration Cloud CIS should deploy Directory Connector on a high-capacity Windows Server (virtual machine or hardware) that is not used to provide other applications and services to the enterprise.

In all cases, it is important to monitor critical on-premises collaboration infrastructure components (Exchange, Active Directory, Directory Connector, Expressway-C, and WebEx Meeting Server); and in cases where servers or virtual machines are failing or where CPU and/or memory usage regularly reach critical levels, consider adding more resources and distributing the load.

Design Considerations for Cloud and Hybrid Services Mobility

Consider the following design requirements and recommendations when enabling and deploying cloud and hybrid services:

- Cisco Directory Connector software must be installed on a Microsoft Windows Server that is a member of the enterprise Active Directory domain, using an account with domain administration privileges.
- Cisco Directory Connector should not be installed on a Window Servers with Active Directory Domain Service (AD DS) or Active Directory Lightweight Directory Services (AD LDS) enabled.
- Integration of Cisco Directory Connector and enterprise Active Directory is recommended for Cisco Spark Hybrid Services to authenticate.
- Cisco Expressway X8.7.1 or later version is required for Cisco Spark Hybrid Services.
- The number of users enabled for calendar service, the size of individual user Exchange calendars, and the rate at which @spark and @webex are used, will determine the amount of increased load on the Exchange server when Cisco Spark Calendar Services are enabled. Create and apply a throttling policy to the Exchange impersonation account to reduce the impact of the Calendar Connector and calendar services on the enterprise Exchange environment.
- A user must be enabled for Call Service Aware in order to be enabled for the Call Service Connect feature.
- Unified CM AXL Web Service and CTIManager services are required for Cisco Spark Call Service and should be enabled on at least two Unified CM nodes to provide high availability of these services.

For more information about the Cisco Collaboration Cloud and Cisco Spark Hybrid Services, refer to the Cisco Spark information available at https://collaborationhelp.cisco.com/article/en-us/nkg4mud.

For information on deploying Cisco Spark Hybrid Services, refer to the latest version of the *Preferred Architecture for Cisco Spark Hybrid Services, CVD*, available at https://www.cisco.com/go/pa.

Mobility Beyond the Enterprise

With Cisco's mobile collaboration solutions, mobility users can handle calls to their enterprise directory number, not only on their desk phone, but also on one or more remote phones. Mobility users can also make calls from a remote phone as if they are dialing inside the enterprise. In addition, mobility users can take advantage of enterprise features such as hold, transfer, and conference as well as enterprise applications such as voicemail, conferencing, and presence on their mobile phones. This ensures continued productivity for users even when they are traveling outside the organization.

Further, with dual-mode phones that provide connectivity to the mobile voice and data provider network as well as the 802.11 WLAN, users not only have the ability to leverage enterprise applications while away from the enterprise, but they can also leverage the enterprise telephony infrastructure when inside the enterprise or remotely attached to the enterprise network to make and receive calls without incurring mobile voice network per-minute charges.

The fixed mobile convergence (FMC) mobility functionality delivered within the Cisco Unified Mobility solution is provided through Cisco Unified CM and can be used in conjunction with Cisco mobile clients and devices such as Cisco Jabber.

Cisco Unified Mobility provides the following mobility application functionality:

- Single Number Reach (SNR)

  Single Number Reach provides users with the ability to be reached at a single enterprise phone number that rings on both their IP desk phone and their mobile phone simultaneously. SNR users can pick up an incoming call on either their desk or mobile phones and at any point can move the in-progress call from one of these phones to the other without interruption.

- Mid-Call Features

  Mid-call features allow a user to invoke hold, resume, transfer, conferencing, and directed call park features from their mobile phone during in-progress mobility calls. These features are invoked from the mobile phone keypad and take advantage of enterprise media resources such as music on hold and conference bridges.

- Single Enterprise Voicemail Box

  Single Enterprise Voicemail box provides mobile voicemail avoidance capabilities and ensures that any unanswered calls made to the user's enterprise number and extended to the user's mobile phone will end up in the enterprise voicemail system rather than in a mobile voicemail system. This provides a single consolidated voicemail box for all business calls and eliminates the need for users to check multiple voicemail systems for messages.

- Mobile Voice Access and Enterprise Feature Access two-stage dialing

  Mobile Voice Access and Enterprise Feature Access two-stage dialing provide mobile users with the ability to make calls from their mobile phone as if they were calling from their enterprise IP desk phone. These features provide a cost savings in terms of toll charges for long distance or international calls as well as calls to internal non-DID extensions on the system that would not normally be reachable from outside the enterprise. These two-stage dialing features also provide the enterprise with an easy way to track phone calls made by users via a uniform and centrally located set of call detail records. Furthermore, these features provide the ability to mask a user's mobile phone number when sending outbound caller ID. Instead, the user's enterprise number is sent as caller ID. This ensures that returned calls to the user are made to the enterprise number, thus resulting in enterprise call anchoring.

Cisco mobile clients and devices provide the ability to attach to both the mobile provider network and 802.11 wireless networks for voice and data connectivity. This enables users to leverage both enterprise call control and in some cases mobile network call control from a single device. By leveraging the enterprise telephony infrastructure for making and receiving calls whenever possible and, in the case of dual-mode phones, falling back to the mobile voice network only when enterprise connectivity is unavailable, mobile clients and devices can help reduce telephony costs. Dual-mode phones and the clients that run on them also provide a handoff mechanism so that in-progress voice calls can be moved easily between the WLAN and mobile voice interfaces as a user moves out of the enterprise.

In addition to enabling mobile devices to make voice or video calls over IP via 802.11 WLAN or mobile data networks, Cisco mobile clients enable automated enterprise dialing using the Dial via Office feature. Dial via Office calls are set up using SIP signaling over the IP network, while the media path is over the mobile voice network and the PSTN. Cisco mobile clients and devices also provide other unified communications services such as corporate directory access, presence and instant messaging (IM). These devices and clients enable mobile users to remain productive whether inside or outside the enterprise by providing access to collaboration applications while at the same time enabling users to make and receive enterprise calls from their mobile devices, whether outside the enterprise over public or private WiFi hot spots or the mobile data network, or inside the enterprise and over the WLAN network.

This section begins with a discussion of Unified Mobility features, functionality, and design and deployment considerations. Given the various benefits of Unified Mobility and the fact that mobile clients and devices can be integrated to take advantage of the features provided, this discussion paves the way for examination of mobile client applications such as Cisco Jabber. This section also includes a discussion of architecture, functionality, and design and deployment implications for the following mobility applications and features:

- Cisco Unified Mobility, page 21-47
- Cisco Mobile Clients and Devices, page 21-76

Cisco Unified Mobility

Cisco Unified Mobility refers to the native mobility functionality within the Cisco Unified CM and includes the Single Number Reach, Mobile Voice Access, and Enterprise Feature Access features.

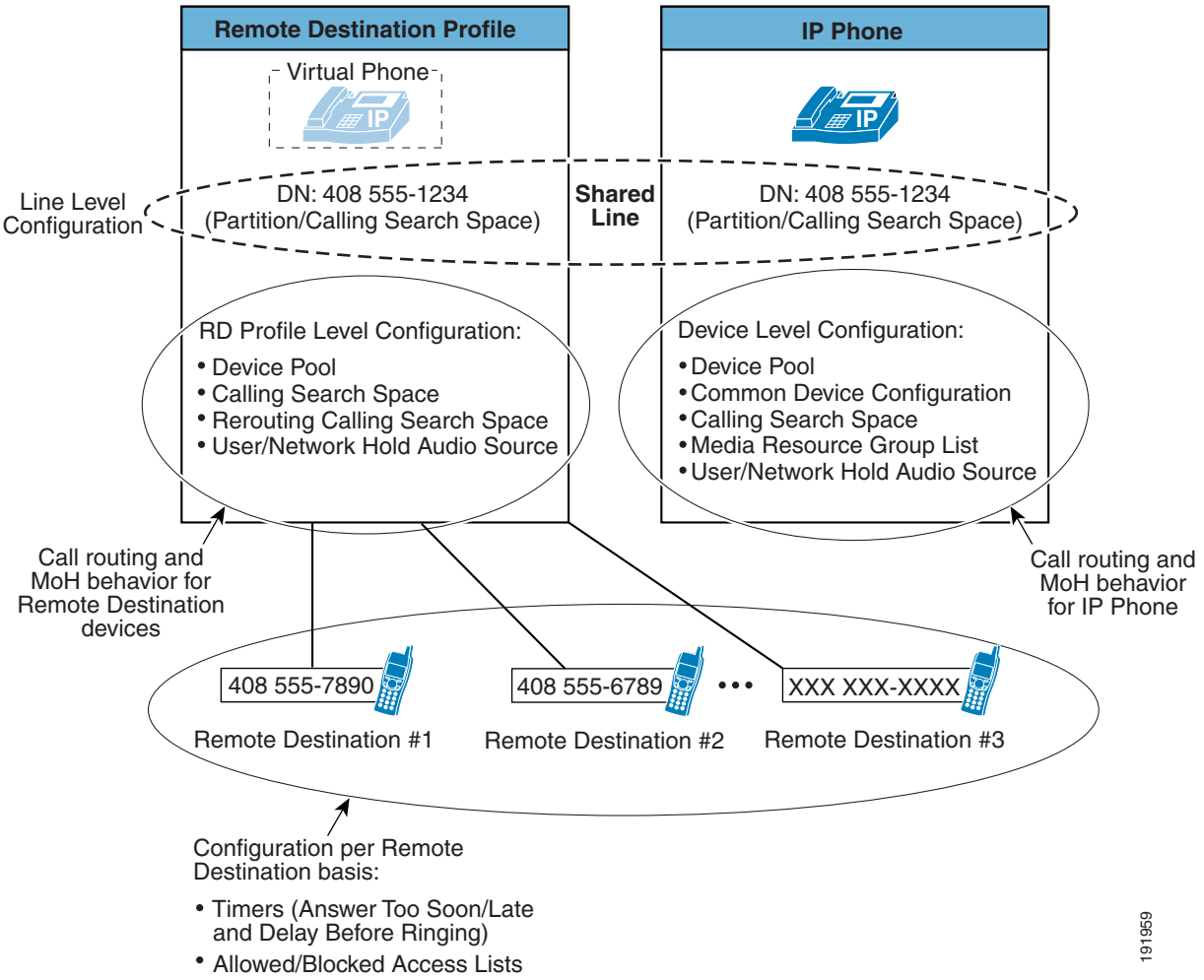
Unified Mobility functionality depends on the appropriate configuration of Unified CM. For this reason, it is important to understand the nature of this configuration as well as the logical components.

Figure 21-17 illustrates the configuration requirements for Unified Mobility. First, as for all users, a mobility user's enterprise phone is configured with appropriate line-level settings such as directory number, partition, and calling search space. In addition, the device-level settings of the enterprise phone include parameters such as device pool, common device configuration, calling search space, media resource group list, and user and network hold audio sources. All of these line and device settings on the user's enterprise phone affect the call routing and music on hold (MoH) behavior for incoming and outgoing calls.

Next, a remote destination profile must be configured for each mobility user in order for them to take advantage of Unified Mobility features. The remote destination profile is configured at the line level with the same directory number, partition, and calling search space as the user's enterprise phone line. This results in a shared line between the remote destination profile and the enterprise phone. The remote destination profile configuration includes device pool, calling search space, rerouting calling search space, and user and network hold audio source parameters. The remote destination profile should be thought of as a virtual phone whose configuration mirrors the user's line-level enterprise phone settings, but whose profile-level configuration combined with the line-level settings determines the call routing

and MoH behavior that the user's remote destination phone will inherit. The user's enterprise directory number, which is shared between the remote destination profile and the enterprise phone, allows calls to that number to be extended to the user's remote destination.

Figure 21-17 Cisco Unified Mobility Configuration Architecture



As further shown in Figure 21-17, a mobility user can have one or more remote destinations configured and associated with their remote destination profile. A remote destination represents a single PSTN phone number where a user can be reached. A user can have up to 10 remote destinations defined. Call routing timers can be configured for each remote destination to adjust the amount of time a call will be extended to a particular remote phone, as well as the amount of time to wait before extending the call and the amount of time that must pass before a call can be answered at the remote phone. Mobility users can also configure filters for each remote destination to allow or deny calls from certain phone numbers to be extended to that remote phone.

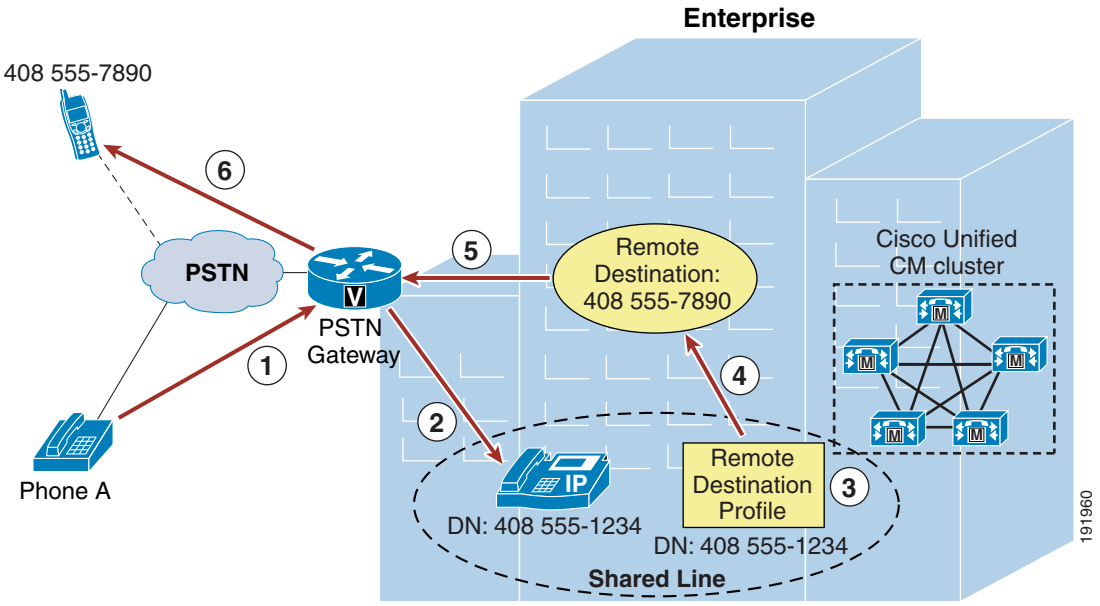
Single Number Reach

The Single Number Reach (SNR) feature allows an incoming call to an enterprise user to be offered to the user's IP desk phone as well as up to 10 configurable remote destinations. Typically a user's remote destination is their mobile or cellular telephone. Once the call is offered to both the desktop and remote destination phone(s), the user can answer at any of those phones. Upon answering the call on one of the remote destination phones or on the IP desk phone, the user has the option to hand off or pick up the call on the other phone.

Single Number Reach Functionality

Figure 21-18 illustrates a basic Single Number Reach call flow. In this example, Phone A on the PSTN calls an SNR user's enterprise directory number (DN) 408-555-1234 (step 1). The call comes into the enterprise PSTN gateway and is extended through Unified CM to the IP phone with DN 408-555-1234 (step 2), and this phone begins to ring. The call is also extended to the user's Remote Destination Profile, which shares the same DN (step 3). In turn, a call is placed to the remote destination associated with the user's remote destination profile (in this case 408-555-7890) (step 4). The outgoing call to the remote destination is routed through the PSTN gateway (step 5). Finally the call rings at the remote destination PSTN phone with number 408 555-7890 (step 6). The call can then be answered at either phone.

Figure 21-18 Single Number Reach



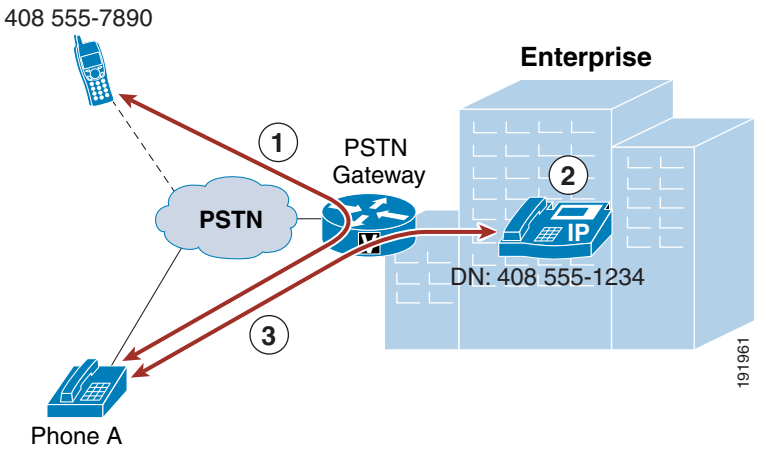
Typically a Single Number Reach user's configured remote destination is their mobile phone on a mobile voice or cellular provider network; however, any destination reachable by means of the PSTN can be configured as a user's remote destination. Furthermore, an SNR user can have up to 10 remote destinations configured, so an incoming call could potentially ring as many as 10 PSTN phones as well as the user's desk phone. Once the call is answered at the desk phone or at a remote destination phone, any other call legs that have been extended to ring additional remote destinations or the desk phone (if not answered at the desk phone) will be cleared. If the incoming call is answered at the remote destination, the voice media path will be hairpinned within the enterprise PSTN gateway utilizing two gateway ports. This utilization must be considered when deploying the SNR feature.

Note In order for Single Number Reach to work as in Figure 21-18, ensure that the user-level Enable Mobility check box under the End User configuration page has been checked and that at least one of the user's configured remote destinations has the Enable Single Number Reach check box checked.

Desk Phone Pickup

As illustrated in Figure 21-19, once a user answers a Single Number Reach call at the remote destination device (step 1: in this case, 408 555-7890), at any point the user can hang up the call at the remote destination and pick it up again at their desk phone by simply pressing the Resume softkey on the desk phone (step 2: at DN 408 555-1234 in this case). The call resumes between the original caller at Phone A and the desk phone (step 3).

Figure 21-19 Desk Phone Pickup



Desk phone pickup can be performed whenever an enterprise-anchored call is in progress at a configured remote destination phone and that phone hangs up the call.

Note An enterprise-anchored call refers to any call that has at least one call leg connected through an enterprise PSTN gateway and that originated either from a remote destination to an enterprise DID or from Single Number Reach, Mobile Voice Access, Enterprise Feature Access, or Intelligent Session Control.

The option to pick up or resume the call at the desk phone is available for a certain amount of time. For this reason, it is good practice for the Single Number Reach user to ensure that the calling phone hangs up before the remote destination phone is hung up. This ensures that the call cannot be resumed at the desk phone by someone else. By default, the call remains available for pickup at the desk phone for 10 seconds after the remote destination phone hangs up; however, this time is configurable and can be set from 0 to 30000 milliseconds on a per-user basis by changing the Maximum Wait Time for Desk Pickup parameter under the End User configuration page. Desk phone pickup can also be performed after invoking the mid-call hold feature at the remote destination phone. However, in these cases, the Maximum Wait Time for Desk Pickup parameter setting has no effect on the amount of time the call will be available for pickup. A call placed on mid-call hold will remain on hold and be available for desk phone pickup until manually resumed at either the remote or desktop phone.

Another method for performing desk phone pickup is to use the mid-call session handoff feature. This mid-call feature is invoked by manually keying *74, the default enterprise feature access code for session handoff, which in turn generates a DTMF sequence back to Unified CM. When this feature is invoked, Unified CM sends a new call to the user's enterprise desk phone. Once this new call is flashing or ringing at the desk phone, the user then must answer the call to complete the session handoff.

The benefit of this desk phone pickup method over other methods (such as hanging up the call at the mobile phone or using the mid-call hold feature) is that the conversation between the user and the far-end phone is maintained throughout the handoff process. Once the *74 sequence has been keyed, the user can continue the conversation because the handoff call is sent to the user's desk phone. When the user answers the call at the desk phone, the call legs are shuffled so that the call leg to the far-end is connected to the new call leg created at the desk phone, thus resulting in an uninterrupted or near-instantaneous cut-through of the audio path. The original call leg at the mobile device is subsequently cleared.

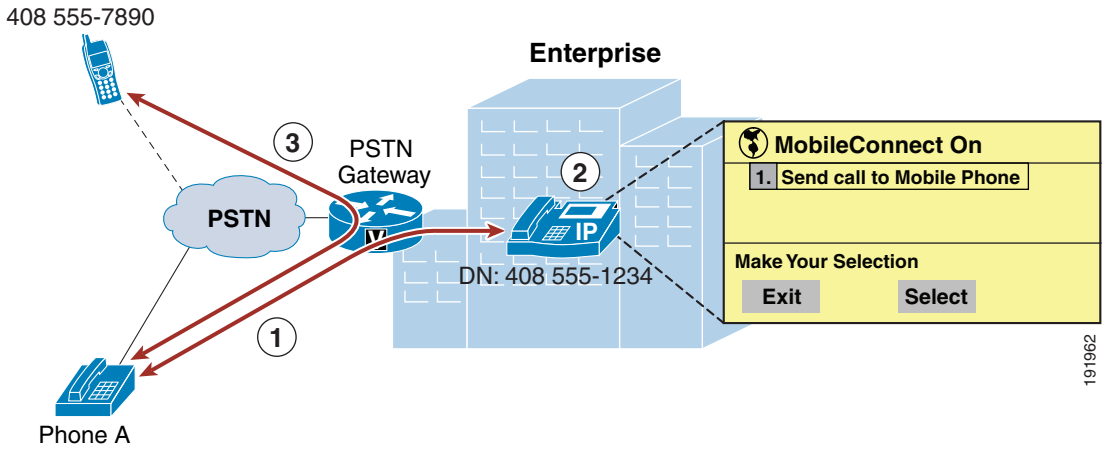
Unlike the hang-up method for invoking desk phone pickup, where the end-user's Maximum Wait Time for Desk Pickup setting determines how long the call will be available for pickup at the desk phone, with session handoff the Session Handoff Alerting Timer service parameter determines the amount of time the call will ring or flash at the desk phone before the handoff call is cleared. The default handoff alerting time is 10 seconds. Further, with session handoff, any call forward settings configured on the desk phone do not get invoked. As a result, the handoff feature does not forward to voicemail or any other call-forward destination. If a call is not answered by the end of Session Handoff Alerting Timer period, then the call is cleared and the Remote In Use state is removed from the user's desk phone line. However, in this scenario the original call is maintained at the mobile phone.

For additional information about session handoff and other mid-call features, see Mid-Call Features, page 21-52.

Remote Destination Phone Pickup

Figure 21-20 illustrates Single Number Reach remote destination phone pickup functionality. Assuming Phone A calls the SNR user's enterprise DN 408 555-1234 and the call is answered at the user's desk phone and is in progress (step 1), the user must push the Mobility softkey. Assuming the SNR feature is enabled for this phone and remote destination pickup is available, the user presses the Select softkey (step 2). A call is generated to the user's remote destination phone (in this case, 408 555-7890), and the remote phone begins to ring. Once the call is answered at the remote phone, the call resumes between Phone A and the SNR user's remote phone with number 408 555-7890 (step 3).

Figure 21-20 Remote Destination Phone Pickup



When a Single Number Reach user has multiple remote destinations configured, each remote destination will ring when the Select softkey is pressed, and the user can answer the desired phone.

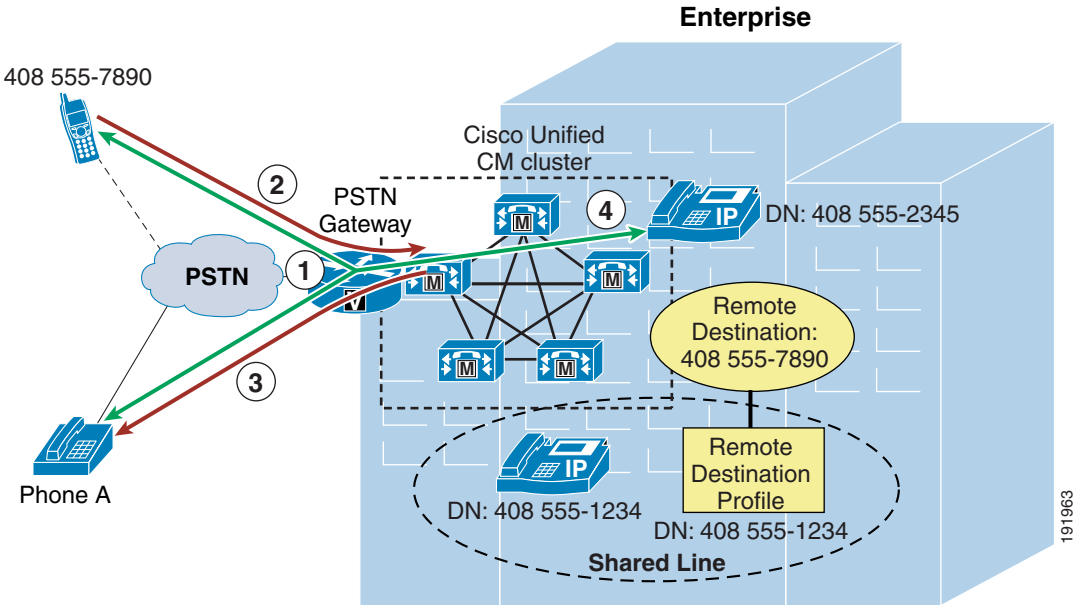
Note In order for remote destination phone pickup to work as in Figure 21-20, ensure that at least one of the user's configured remote destinations has the Mobile Phone check box checked. In addition, the Mobility softkey must be configured for all mobility users by adding the softkey to each user's associated desk phone softkey template. Failure to check the Mobile Phone check box and to make the Mobility softkey available to mobility users will prevent the use of remote destination phone pickup functionality.

Note Cisco TelePresence System C, EX, MX, SX, and TX Series video endpoints do not support remote destination pickup as described above. These endpoints do not expose a mobility softkey or the "Send call to Mobile Phone" option to the user. Therefore, these endpoints are unable to send in-progress calls to the mobile device using remote destination pickup.

Mid-Call Features

As illustrated in Figure 21-21, once a user answers a Single Number Reach call at the remote destination device (step 1: in this case, 408 555-7890), the user can invoke mid-call features such as hold, resume, transfer, conference, directed call park, and session handoff by sending DTMF digits from the remote destination phone to Unified CM via the enterprise PSTN gateway (step 2). When the mid-call feature hold, transfer, conference, or directed call park is invoked, MoH is forwarded from Unified CM to the held party (step 3: in this case, Phone A). In-progress calls can be transferred to another phone or directed call park number, or additional phones can be conferenced using enterprise conference resources (step 4).

Figure 21-21 Mobility Mid-Call Feature



Mid-call features are invoked at the remote destination phone by a series of DTMF digits forwarded to Unified CM. Once received by Unified CM, these digit sequences are matched to the configured Enterprise Feature Access Codes for Hold, Exclusive Hold, Resume, Transfer, Conference, and Session Handoff, and the appropriate function is performed.

Note To enable the Directed Call Park mid-call feature, you must configure Cisco Unified CM with directed call park numbers and call park retrieval prefixes.

Note In order to perform the transfer, conference, and directed call park mid-call features, a second call leg is generated by the remote destination phone to a system-configured Enterprise Feature Access DID that answers the call, takes user input (including PIN number, mid-call feature access code, and target number), and then creates the required call leg to complete the transfer, conference, or directed call park operation.

With the mid-call session handoff feature, MoH is not forwarded to the far-end because the far-end is never placed on hold. Instead, the original audio path is maintained until the mobile user answers the handoff call at the desk phone. Once the call is answered, the call legs are shuffled at the enterprise gateway and the audio path is maintained.

Mid-call features are invoked by manually keying the feature access codes and entering the appropriate key sequences. Table 21-2 indicates the required key sequences for invoking mid-call features.

Table 21-2 Manual Mid-Call Feature Key Sequences

Mid-Call Feature	Enterprise Feature Access Code (default)	Manual Key Sequence
Hold	*81	Enter: *81
Exclusive Hold	*82	Enter: *82
Resume	*83	Enter: *83
Transfer	*84	1. Enter: *82 (Exclusive Hold) 2. Make new call to Enterprise Feature Access DID. 3. On connect, enter: <*PIN_number*> # *84 # <*Transfer_Target/DN*> # 4. Upon answer by transfer target (for consultive transfer) or upon ringback (for early attended transfer), enter: *84
Directed Call Park	N/A	1. Enter: *82 (Exclusive Hold) 2. Make new call to Enterprise Feature Access DID. 3. On connect, enter: <*PIN_number*> # *84 # <*Directed_Call_Park_Number*> # *84 # **Note** To retrieve a parked call, the user must use Mobile Voice Access or Enterprise Feature Access Two-Stage Dialing to place a call to the directed call park number. When entering the directed call park number to be dialed, it must be prefixed with the appropriate call park retrieval prefix.
Conference	*85	1. Enter: *82 (Exclusive Hold) 2. Make new call to Enterprise Feature Access DID. 3. On connect enter: <*PIN_number*> # *85 # <*Conference_Target/DN*> # 4. Upon answer by conference target, enter: *85
Session Handoff	*74	1. Enter: *74 2. Answer at the desk phone upon ring and/or flash.

Note Media resource allocation for mid-call features such as hold and conference is determined by the Remote Destination Profile configuration or, in the case of dual-mode phones and Unified Mobile Communicator, the device configuration. The media resource group list (MRGL) of the device pool configured for the Remote Destination Profile or the mobile client device is used to allocate a conference bridge for the conferencing mid-call feature. The User Hold Audio Source and Network Hold MoH Audio Source settings of the Remote Destination Profile or the mobile client device, in combination with the media resource group list (MRGL) of the device pool, is used to determine the appropriate MoH stream to be sent to a held device.

Mobile Voicemail Avoidance with Single Enterprise Voicemail Box

An additional consideration with Cisco Unified Mobility Single Number Reach is mobile voicemail avoidance. The single enterprise voicemail box feature ensures that all unanswered enterprise business calls end up at the enterprise voicemail system. This prevents a user from having to check multiple mailboxes (enterprise, mobile, home, and so forth) for calls to their enterprise phone number that are unanswered. This feature provides two methods for avoiding mobile or non-enterprise voicemail:

- Timer Control method — With this method the system relies on a set of timers (one per remote destination) in conjunction with system call-forward timers to ensure that, when and if a call is forwarded to a voicemail system on ring-no-answer, the enterprise voicemail system receives the call.
- User Control method — With this method the system relies on a DTMF confirmation tone from the remote destination when the call is answered to determine if the call was received by the user or a non-enterprise voicemail system.

System settings determine whether the timer control or user control method is used. The method used can be set globally via the Voicemail Selection Policy service parameter or for individual remote destinations via the Single Number Reach Voicemail Policy. By default the system and all remote destinations use the timer control method

Timer Control Mobile Voicemail Avoidance

For this method, the system relies on a set of timers on the Remote Destination configuration page. The purpose of these timers is to ensure that, when and if a call is forwarded to a voicemail system on ring-no-answer, the call is forwarded to the enterprise voicemail system rather than any remote destination voicemail system. These timers in conjunction with other system forward-no-answer timers should be configured to avoid non-enterprise voicemail systems as follows:

- Ensure the system forward-no-answer time is shorter at the desk phone than at the remote destination phones.

  To do so, ensure that the global Forward No Answer Timer field in Unified CM or the No Answer Ring Duration field under the individual phone line is configured with a value that is less than the amount of time a remote destination phone will ring before forwarding to the mobile voicemail system. In addition, the Delay Before Ringing Timer parameter under the Remote Destination configuration page can be used to delay the ringing of the remote destination phone in order to further lengthen the amount of time that must pass before a remote destination phone will forward to its own mobile voicemail box. However, when adjusting the Delay Before Ringing Timer parameter, take care to ensure that the global Unified CM Forward No Answer Timer (or the line-level No Answer Ringer Duration field) is set sufficiently high enough so that the mobility user has time to answer the call on the remote destination phone. The Delay Before Ringing Timer parameter can be set for each remote destination and is set to 4,000 milliseconds by default.
- Ensure that the remote destination device stops ringing before the incoming call is forwarded to the mobile voicemail system.

  You can accomplish this with the Answer Too Soon and Answer Too Late timers for each remote destination. First the Answer Too Soon Timer parameter under the Remote Destination configuration page should be configured with a value that is more than the amount of time it takes a call extended to a powered-off or out-of-range mobile phone to be forwarded to the mobile voicemail system. By default this timer is set 1,500 milliseconds (or 1.5 seconds). If the call is answered before the Answer Too Soon Timer expires, the system will disconnect the call leg to the remote destination. This ensures that calls forwarded immediately to the mobile voicemail system will not be connected, but those answered by the user after ring-in are connected.

Next configure the Answer Too Late Timer parameter under the Remote Destination configuration page with a value that is less than the amount of time that a remote destination phone will ring before forwarding to its voicemail box. By default this timer is set to 19,00 milliseconds (or 19 seconds). If the call is not answered before this timer expires, the system will disconnect the call leg to the remote destination. This ensures that the remote destination phone stops ringing before the call is forwarded to the mobile voicemail system.

Note Incoming calls to a remote destination that are manually diverted by the mobility user can end up in the mobile voicemail box if the manual diversion occurs after the Answer Too Soon timer has expired. To prevent this from happening, mobility users should be configured for the user control method or advised to ignore or silence the ringing of incoming calls they wish to divert to voicemail. This will ensure that unanswered calls always end up in the enterprise voicemail system.

Note In most deployment scenarios, the default Delay Before Ringing Timer, Answer Too Late Timer, and Answer Too Soon Timer values are sufficient and do not need to be changed.

User Control Mobile Voicemail Avoidance

For this method, the system relies on DTMF confirmation tone from the remote destination when the call is answered. If a DTMF tone is received by the system, then the system knows that the user answered the call and pressed a key to generate the DTMF tone. On the other hand, if the DTMF tone is not received by the system, the system assumes the call leg was answered by a non-enterprise voicemail system and it disconnects the call leg.

When the user control method is enabled, on answer the end user will hear an audio prompt requesting that they press a key pad button to generate a DTMF tone. By default the audio prompt is played to the user one second after the call is answered. The user may not hear the audio prompt if they press the keypad to generate a DTMF tone immediately upon answering. The audio prompt is played only on the remote destination call leg and therefore the far-end party will not hear this prompt. Once the audio prompt is played to the user, by default the system will wait 5 seconds to receive the DTMF tone. If the tone is not received, the system disconnects the call leg but continues to ring the user's other configured devices until the call is answered by the user or forwarded to the enterprise voicemail system.

Note The user control mobile voicemail avoidance method is completely dependent on successful relay of the DTMF tone from the remote destination on the mobile voice network or PSTN all the way to Unified CM. The DTMF tone must be sent out-of-band to Unified CM. If DTMF relay is not properly configured on the network and system, DTMF will not be received and all call legs to remote destinations relying on the user control method will be disconnected. The system administrator should ensure proper DTMF interoperation and relay across the enterprise telephony network prior to enabling the user control method. If DTMF cannot be effectively relayed from the PSTN to Unified CM, then the timer control mobile voicemail avoidance method should be used instead.

Enabling and Disabling Single Number Reach

The Single Number Reach (SNR) feature can be enabled or disabled by using one of the following methods:

- Cisco Unified CM Administration or Cisco Unified CM Self Care Portal for end users

  An administrator or user unchecks the Enable Single Number Reach box to disable, or checks the Enable Single Number Reach box to enable, the feature. This is done per remote destination.

- Mobile Voice Access or Enterprise Feature Access

  A Mobility-enabled user dials into the Mobile Voice Access or Enterprise Feature Access DID and, after entering appropriate credentials, enters the digit 2 to enable or 3 to disable. With Mobile Voice Access, the user is prompted to enable or disable SNR for a single remote destination or all of their remote destinations. With Enterprise Feature Access, the user can enable or disable SNR only for the remote destination device from which they are calling.

- Desk phone Mobility softkey or icon

  The user presses the Mobility softkey when the phone is in the on-hook state and selects either Enable Mobile Connect or Disable Mobile Connect. On some phone models the user touches the mobility icon and then selects **Off** to disable Single Number Reach. Alternatively, the user can select **Ring only this phone**. To enable Single Number Reach again the user selects **Ring all devices**. With any of these methods, Single Number Reach is enabled or disabled for all of the user's remote destinations.

Note The dialog box that appears when the Mobility softkey is pressed as described above uses the old feature name, Mobile Connect, rather than the new feature name, Single Number Reach. The feature and enable/disable functionality are the same.

Access Lists for Allowing or Blocking Single Number Reach Calls

Access lists can be configured within Cisco Unified CM and associated to a remote destination. Access lists are used to allow or block inbound calls (based on incoming caller ID) from being extended to a mobility-enabled user's remote destinations. Furthermore, these access lists are invoked based on the time of day.

Access lists are configured for mobility-enabled users as either blocked or allowed. Access lists contain one or more members or filters consisting of a specific number or number mask, and the filters are compared against the incoming caller ID of the calling party. In addition to containing specific number strings or number masks for matching caller ID, access lists can also contain a filter for incoming calls where the caller ID is not available or is set to private. A blocked access list contains an implicit "allow all" at the end of the list so that calls from any numbers entered in the access list will be blocked but calls from all other numbers will be allowed. An allowed access list contains an implicit "deny all" at the end of the list so that calls from any numbers entered in the access list will be allowed but calls from all other numbers will be blocked.

Once configured access lists are associated with a configured Ring Schedule under the Remote Destination configuration screen, the configured Ring Schedule in combination with the selected access list provides time-of-day call filtering for Single Number Reach calls on a per-remote-destination basis. Access lists and Ring Schedules can be configured and associated to a remote destination by an administrator using the Cisco Unified CM Administration interface or by an end user using the Cisco Unified CM Self Care Portal.

Single Number Reach Architecture

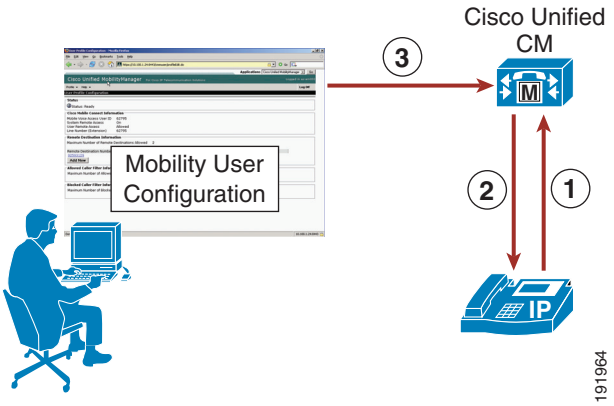
The architecture of the Single Number Reach (SNR) feature is as important to understand as its functionality. Figure 21-22 depicts the message flows and architecture required for SNR. The following sequence of interactions and events can occur between Unified CM, the SNR user, and the SNR user's desk phone:

1. The SNR phone user who wishes to either enable or disable the SNR feature or to pick up an in-progress call on their remote destination phone pushes the Mobility softkey on their desk phone (see step 1 in Figure 21-22).
2. Unified CM returns the SNR status (On or Off) and offers the user the ability to select the Send Call to Mobile Phone option when the phone is in the Connected state, or it offers the user the ability to enable or disable the Mobile Connect status when the phone is in the On Hook state (see step 2 in Figure 21-22).
3. Single Number Reach users can use the Unified CM Self Care Portal to configure their own mobility settings via the web-based configuration pages at

   https://<*Unified-CM_Server_IP_Address*>/ucmuser/

   where <*Unified-CM_Server_IP_Address*> is the IP address of the Unified CM publisher server (see step 3 in Figure 21-22).

Figure 21-22 Single Number Reach Architecture



High Availability for Single Number Reach

The Single Number Reach feature relies on the following components:

- Unified CM servers
- PSTN gateway

Each component must be redundant or resilient in order for Single Number Reach to continue functioning fully during various failure scenarios.

Unified CM Server Redundancy

The Unified CM server is required for the Single Number Reach feature. Unified CM server failures are non-disruptive to SNR functionality, assuming phone and gateway registrations are made redundant using Unified CM Groups.

In order for SNR users to use the Unified CM Self Care Portal web interface to configure their mobility settings (remote destinations and access lists), the Unified CM publisher server must be available. If the publisher is down, users will not be able to change mobility settings. Likewise, administrators will be unable to make mobility configuration changes to Unified CM; however, existing mobility configurations and functionality will continue. Finally, changes to SNR status must be written by the system on the Unified CM publisher server; if the Unified CM publisher is unavailable, then enabling or disabling SNR will not be possible.

PSTN Gateway Redundancy

Because the Single Number Reach feature relies on the ability to extend additional call legs to the PSTN to reach the SNR users' remote destination phones, PSTN gateway redundancy is important. Should a PSTN gateway fail or be out of capacity, the SNR call cannot complete. Typically, enterprise IP telephony dial plans provide redundancy for PSTN access by providing physical gateway redundancy and call re-routing capabilities as well as enough capacity to handle expected call activity. Assuming that Unified CM has been configured with sufficient capacity, multiple gateways, and route group and route list constructs for call routing resiliency, the SNR feature can rely on this redundancy for uninterrupted functionality.

Mobile Voice Access and Enterprise Feature Access

Mobile Voice Access (also referred to as System Remote Access) and Enterprise Feature Access two-stage dialing are features built on top of the Single Number Reach application. Both features allow a mobility-enabled user who is outside the enterprise to make a call as though they are directly connected to Unified CM. This functionality is commonly referred to as Direct Inward System Access (DISA) in traditional telephony environments. These features benefit the enterprise by limiting toll charges and consolidating phone billing directly to the enterprise rather than billing to each mobile user. In addition, these features allow the users to mask their mobile phone or remote destination numbers when sending outbound caller ID. Instead, the user's enterprise directory number is sent as caller ID. This ensures that returned calls to the user are made to the enterprise number, thus resulting in enterprise call anchoring. These features also enable mobile users to dial internal extensions or non-DID enterprise numbers that would not normally be reachable from outside the enterprise.

Mobile Voice Access is accessed by calling a system-configured DID number that is answered and handled by an H.323 or SIP VoiceXML (VXML) gateway. The VoiceXML gateway plays interactive voice response (IVR) prompts to the Mobile Voice Access user, requesting user authentication and input of a number to be dialed via the user phone keypad.

Enterprise Feature Access functionality includes the previously discussed mid-call transfer and conference features as well as two-stage dialing functionality. Two-stage dialing works the same way as Mobile Voice Access, but without the IVR prompts. The system-configured Enterprise Feature Access DID is answered by Unified CM. The user then uses the phone keypad or Smart Phone softkeys to input authentication and the number to be dialed. These inputs are received without prompts.

With both the Mobile Voice Access and Enterprise Feature Access two-stage dialing features, once the call to the input number is connected, users can invoke mid-call features or pick up the call on their desk phones just as with a Single Number Reach call. This is possible because the call is anchored at the enterprise gateway.

Mobile Voice Access IVR VoiceXML Gateway URL

The Mobile Voice Access feature requires the Unified CM VoiceXML application to reside on the H.323 or SIP gateway. The URL used to load this application is:

http://<*Unified-CM-Publisher_IP-Address*>:8080/ccmivr/pages/IVRMainpage.vxml

where <*Unified-CM-Publisher_IP-Address*> is the IP address of the Unified CM publisher node.

Mobile Voice Access Functionality

Figure 21-23 illustrates a Mobile Voice Access call flow. In this example, the Mobile Voice Access user on PSTN phone 408 555-7890 dials the Mobile Voice Access enterprise DID DN 408-555-2345 (step 1).

The call comes into the enterprise PSTN H.323 or SIP gateway, which also serves as the VoiceXML gateway (step 2).

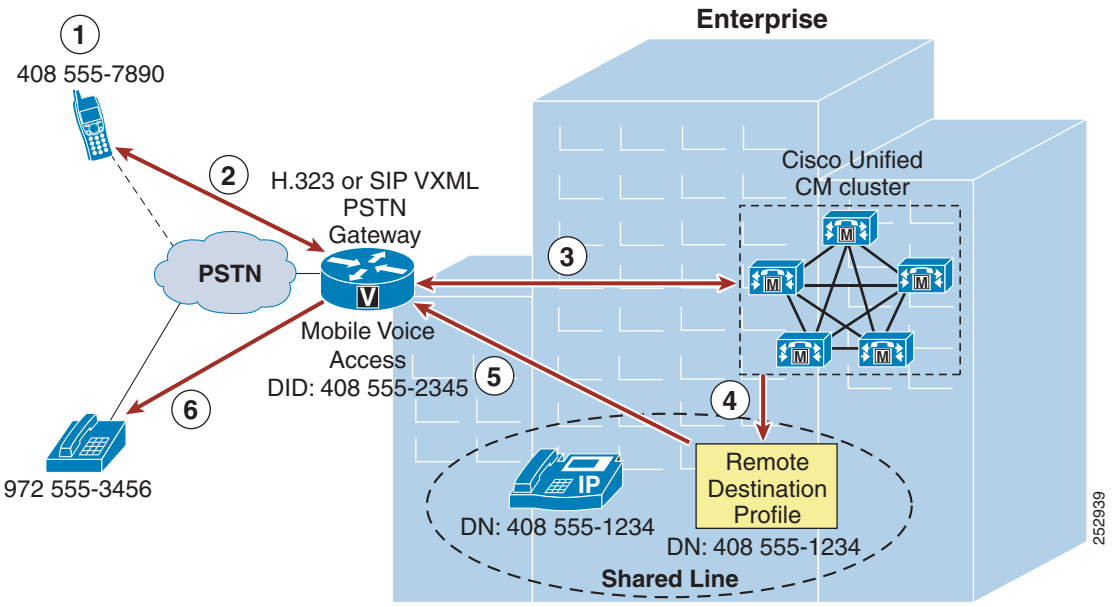
Note Native VoiceXML support is not available with Cisco IOS XE; therefore, the Cisco 4000 Series Integrated Services Router (ISR) cannot be deployed as a VoiceXML gateway for Mobile Voice Access. Instead a Cisco IOS gateway supporting native VXML must be used.

The user is prompted via IVR to enter their numeric user ID (followed by the # sign), PIN number (followed by the # sign), and then a 1 to make a Mobile Voice Access call, followed by the phone number they wish to reach. In this case, the user enters 9 1 972 555 3456 as the number they wish to reach (followed by the # sign).

Note If the PSTN phone from which the Mobile Voice Access user is calling is configured as a Single Number Reach remote destination for that user and the incoming caller ID can be matched against this remote destination by Unified CM, the user does not have to enter their numeric user ID. Instead they will be prompted to enter just the PIN number.

In the meantime, Unified CM has forwarded IVR prompts to the gateway, the gateway has played these prompts to the user, and the gateway has collected user input including the numeric ID and PIN number of the user. This information is forwarded to Unified CM for authentication and to generate the call to 9 1 972 555 3456 (step 3). After authenticating the user and receiving the number to be dialed, Unified CM generates a call via the user's Remote Destination Profile (step 4). The outbound call to 972 555-3456 is routed via the PSTN gateway (step 5). Finally, the call rings at the PSTN destination phone with number 972 555-3456 (step 6).

Figure 21-23 Mobile Voice Access



Note In order for Mobile Voice Access to work as in Figure 21-23, ensure that the system-wide Enable Mobile Voice Access service parameter is set to True and that the per-user Enable Mobile Voice Access check box on the End User configuration page is also checked.

Note The Mobile Voice Access feature relies on the Cisco Unified Mobile Voice Access Service, which must be activated manually from the Unified CM Serviceability configuration page. This service can be activated on the publisher node only.

Note If the PSTN gateway is a Cisco 4000 Series ISR which does not support native VoiceXML, the VoiceVXML functionality required for Mobile Voice Access must be offloaded to an H.323 Cisco IOS gateway with native VoiceXML support using the hairpinning method of deployment as described in the next section.

Mobile Voice Access Using Hairpinning

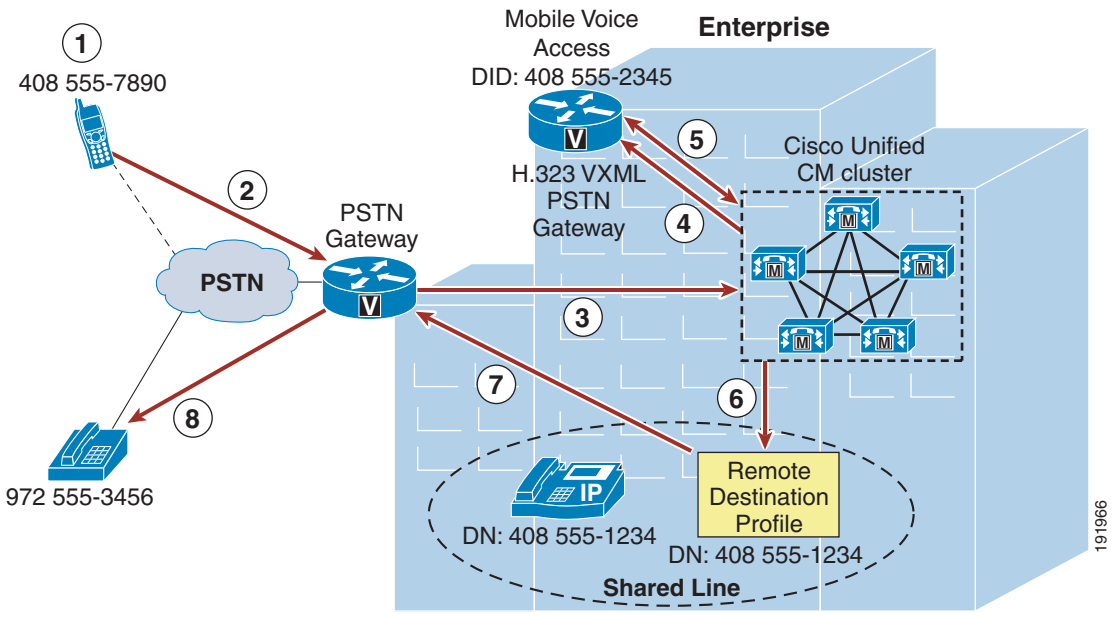
In deployments where the enterprise PSTN gateways are not using H.323 or SIP, Mobile Voice Access functionality can still be provided using hairpinning on a separate gateway running H.323. Mobile Voice Access using hairpinning relies on off-loading the VoiceXML functionality to a separate H.323 gateway. Figure 21-24 illustrates a Mobile Voice Access call flow using hairpinning. In this example, just as in the previous example, the Mobile Voice Access user on PSTN phone 408 555-7890 dials the Mobile Voice Access enterprise DID DN 408-555-2345 (step 1). The call comes into the enterprise PSTN gateway (step 2) and is forwarded to Unified CM for call handling (step 3). Unified CM next routes the inbound call to the H.323 VoiceXML gateway (step 4). The user is then prompted by IVR to enter their

numeric user ID, PIN, and then a 1 to make a Mobile Voice Access call, followed by the phone number they wish to reach. Again the user enters 9 1 972 555 3456 as the number they wish to reach (followed by the # sign).

Note When using Mobile Voice Access with hairpinning, users calling into the system will not be identified automatically by their caller ID. Instead, users will have to key in their remote destination number manually prior to entering their PIN. The reason the user is not automatically identified is that, for hairpinning deployments, the PSTN gateway must first route the call to Unified CM to reach the hairpinned Mobile Voice Access gateway. Because the call is routed to Unified CM first, the conversion of the calling number from a mobile number to an enterprise directory number occurs prior to the call being handled by the Mobile Voice Access gateway. This results in the Mobile Voice Access gateway being unable to match the calling number with a configured remote destination, and therefore the system prompts the user to enter their remote destination number. This is unique to hairpinning deployments; with normal Mobile Voice Access flows, the PSTN gateway does not have to route the call to Unified CM first in order to access Mobile Voice Access because the functionality is available on the local gateway.

In the meantime, the H.323 VoiceXML gateway collects and forwards the user input to Unified CM and then plays the forwarded IVR prompts to the PSTN gateway and the Mobile Voice Access user. Unified CM in turn receives user input, authenticates the user, and forwards appropriate IVR prompts to the H.323 VoiceXML gateway based on user input (step 5). After receiving the number to be dialed, Unified CM generates a call using the user's Remote Destination Profile (step 6). The outbound call to 972 555-3456 is routed through the PSTN gateway (step 7). Finally, the call rings at the PSTN destination phone with number 972 555-3456 (step 8).

Figure 21-24 Mobile Voice Access Using Hairpinning



Note When deploying Mobile Voice Access in hairpinning mode, Cisco recommends configuring the Mobile Voice Access DID at the PSTN gateway and the Mobile Voice Access Directory Number within Cisco Unified CM (under **Media Resources > Mobile Voice Access**) as different numbers. A translation pattern within Unified CM can then be used to translate the called number of the Mobile Voice Access DID to the configured Mobile Voice Access directory number. Because the Mobile Voice Access directory number configured within Unified CM is visible to the administrator only, translation between the DID and directory number will be invisible to the end user and there will be no change in end-user dialing behavior. This is recommended in order to prevent mobility call routing issues in multi-cluster environments. This recommendation does not apply to Mobile Voice Access in non-hairpinning mode.

Note Mobile Voice Access in hairpinning mode is supported only with H.323 VXML gateways.

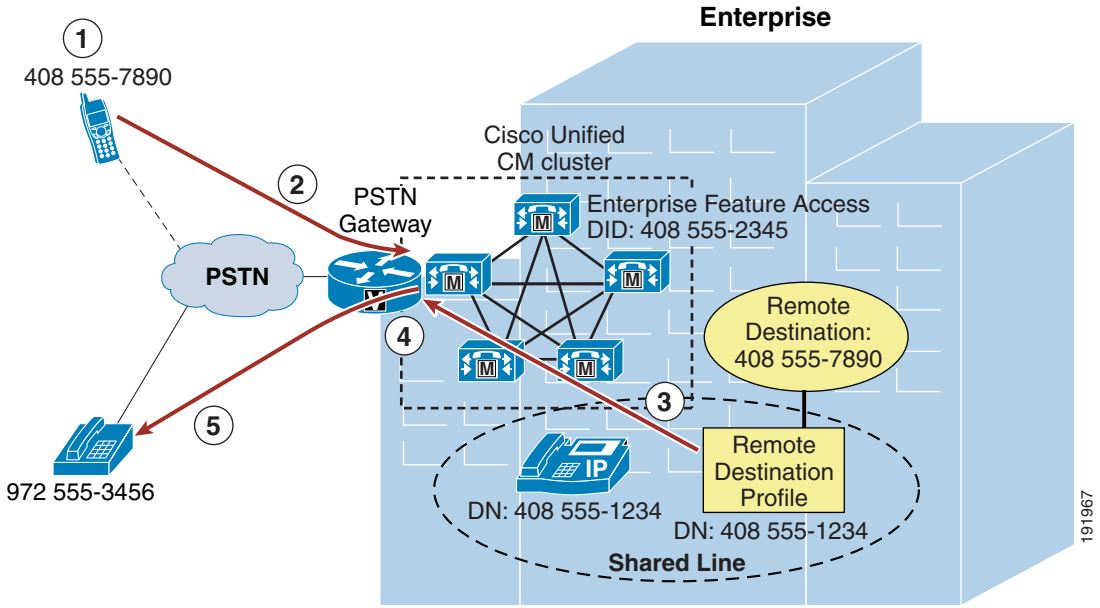
Enterprise Feature Access with Two-Stage Dialing Functionality

Figure 21-25 illustrates the call flow for Enterprise Feature Access two-stage dialing. In this example, the mobility user at remote destination phone 408 555-7890 dials the Enterprise Feature Access DID 408 555-2345 (step 1). Once the call is connected, the remote destination phone is used to send DTMF digits to Unified CM via the PSTN gateway, beginning with the user's PIN (followed by the # sign) which is authenticated with Unified CM. Next a 1 (followed by the # sign) is sent to indicate a two-stage dialed call is being attempted, followed by the phone number the user wishes to reach. In this case the user enters 9 1 972 555 3456 as the destination number (step 2).

Note Unlike with Mobile Voice Access, Enterprise Feature Access requires that all two-stage dialed calls must originate from a phone that has been configured as a remote destination in order to match the caller ID and PIN against the end-user account. There is no provision within Enterprise Feature Access in which the mobility user can enter their remote destination number or ID to identify themselves to the system. Identity can be established only via the combination of incoming caller ID and entered PIN.

Next the outgoing call is originated via the user's remote destination profile (step 3), and the call to PSTN number 972 555-3456 is routed via the enterprise PSTN gateway (step 4). Finally, the call rings the PSTN phone (step 5: in this case, 972 555-3456). As with Mobile Voice Access, the voice media path of each Enterprise Feature Access two-stage dialed call is hairpinned within the enterprise PSTN gateway utilizing two gateway ports.

Figure 21-25 Enterprise Feature Access Two-Stage Dialing Feature



Note In order for Enterprise Feature Access two-stage dialing to work as in Figure 21-25, ensure that the system-wide Enable Enterprise Feature Access service parameter is set to True.

Desk and Remote Destination Phone Pickup

Because Mobile Voice Access and Enterprise Feature Access functionality is tightly integrated with the Single Number Reach feature, once a Mobile Voice Access or Enterprise Feature Access two-stage dialed call has been established, the user does have the option of using Single Number Reach functionality to pick up the in-progress call on their desk phone by simply hanging up the call on the originating phone and pushing the Resume softkey on their desk phone or by using the mid-call hold feature. In turn, the call can then be picked up on the user's configured remote destination phone by pressing the Mobility softkey and selecting Send Call to Mobile Phone.

Enabling and Disabling Single Number Reach

In addition to providing users of Mobile Voice Access and Enterprise Feature Access with the ability to make calls from the PSTN as though they are within the enterprise, the functionality provided by Mobile Voice Access on the H.323 or SIP VoiceXML gateway and provided by Enterprise Feature Access also gives users the ability to remotely enable and disable their Single Number Reach functionality for each remote destination via their phone keypad. Rather than entering a 1 to make a call, users enter a 2 to turn the Single Number Reach feature on and a 3 to turn the Single Number Reach feature off.

If a user has more than one remote destination configured when using Mobile Voice Access, they are prompted to key in the remote destination phone number for which they wish to enable or disable the Single Number Reach feature. When using Enterprise Feature Access, a user can enable or disable Single Number Reach only for the remote destination phone from which they are calling.

Note When the Enable Mobile Voice Access service parameter is set to False, resulting in an inability to make two-stage dialed calls, Mobile Voice Access still provides users with the ability to enable and disable Single Number Reach remotely. As long as the Mobile Voice Access Directory Number has been configured on the system, the user's account has been enabled for Mobile Voice Access, and the Cisco Unified Mobile Voice Access service is running on the publisher, an authorized calling user can still enable or disable Single Number Reach.

Mobile Voice Access and Enterprise Feature Access Number Blocking

Administrators might want to prevent users of Mobile Voice Access and Enterprise Feature Access two-stage dialing from dialing certain numbers when using these features. In order to restrict or block calls to certain numbers when using these features for off-net calls, a comma-separated list of those numbers can be configured in the System Remote Access Blocked Numbers service parameter field. Once this parameter is configured with blocked numbers, those numbers will not be reachable from a user's remote destination phone when using Mobile Voice Access or Enterprise Feature Access features. Numbers that administrators might want to block can include emergency numbers such as 911. When configuring blocked numbers, ensure they are configured as they would be dialed by an enterprise user, with appropriate prefixes or steering digits. For example, if an emergency number is to be blocked and the emergency number is dialed by system users as 9911, then the number configured in the System Remote Access Blocked Numbers field should be 9911.

Access Numbers for Mobile Voice Access

While the Unified CM system allows the configuration of only a single Mobile Voice Access Directory Number, this does not preclude the use of multiple externally facing numbers that can access these internally configured numbers. For example, consider a system deployed in the US in New York with a remote site in San Jose as well as an overseas site in London. Even though the system may have the Mobile Voice Access directory number configured as 555-1234, the gateways at each location can be configured to map a local or toll-free DID number to this Mobile Voice Access directory number. For example, the gateway in New York may have DIDs of +1 212 555 1234 and +1 800 555 1234, which both map to the Mobile Voice Access number, while the gateway in San Jose has a DID of +1 408 666 5678 and the gateway in London has a DID of +44 208 777 0987, which also map to the Mobile Voice Access number of the system.

By acquiring multiple local or toll-free DID numbers, system administrators can ensure that Mobile Voice Access two-stage dialed calls will always originate as a call into the system that is either local or toll-free, thus providing further reductions in telephony costs.

Remote Destination Configuration and Caller ID Matching

When authenticating users for Mobile Voice Access and Enterprise Feature Access two-stage dialing functionality as well as the DTMF-based mid-call features Transfer and Conference, the caller ID of the calling remote destination phone is matched against all remote destinations configured within the system. Matching of this caller ID depends on a number of factors, including how the remote destination numbers are configured, whether digit prefixing is required to include PSTN steering digits on the system, and whether the Matching Caller ID with Remote Destination parameter is set to Partial or Complete Match. In all cases, the requirement is to be able to uniquely identify each mobility user based on the their remote destination number or numbers. For this reason, it is critical not only that remote destination numbers be configured uniquely within the system, but also that inbound caller ID matching (whether using complete or partial matching) must always uniquely correspond to a single remote destination. If a single or unique match is not found, caller ID matching will fail.

To control the nature of this matching, consider the following two approaches.

Using Complete Caller ID Matching

With this approach, remote destination numbers are configured exactly as the caller ID would be presented from the PSTN. For example, if the caller ID from the PSTN for a remote destination phone is presented to the system as 4085557890, then this number should be configured on the Remote Destination configuration page.

In order to route Single Number Reach calls appropriately to this remote destination, it is necessary to configure the dial plan to use either +E.164 dialing methods or a digit prefix mechanism to prefix necessary PSTN access codes and other required digits. For example, if you are not using a global +E.164 dial plan and assuming a 9 or other PSTN steering digits or country codes are required to reach the PSTN when dialing calls from the enterprise, then digit prefixing must be configured to add the appropriate PSTN steering digit and country code to the beginning of the configured remote destination number. Digit prefixing should be facilitated by using translation patterns, route patterns, or route list constructs within the Unified CM system. When using this complete match approach and a digit prefixing method, the Matching Caller ID with Remote Destination parameter should be left at the default setting of **Complete Match**.

Application Dial Rules may also be used to provide digit prefixing in these scenarios. However, it is worth noting that Application Dial Rules are applied based on called digit-string length and cannot be partitioned, meaning that they are applied globally across the system. This severely limits the use of Application Dial Rules, especially in scenarios where multiple dialing domains (for example, different countries) need to be supported on a single Unified CM cluster.

Note Not only are Application Dial Rules applied to Single Number Reach, Mobile Voice Access, and Enterprise Feature Access calls, but they are also applied to calls made with Cisco WebDialer, Cisco Unified CM Assistant, and Cisco Jabber applications. For this reason, exercise care when configuring these rules to ensure that dialing behavior across all applications is as expected.

The recommended dial plan approach is always to globalize the caller ID to +E.164 on ingress from the PSTN and always to configure remote destinations as +E.164. This will guarantee that the caller ID from the PSTN (after normalization) will always provide a unique match when compared against all configured remote destinations. Combined with a dial plan supporting +E.164 dialing, this eliminates the need for digit prefixing and ensures unique identification of remote destination users and numbers even when supporting multiple international numbering plans. Because the recommended dial plan approach is to globalize the caller ID on ingress and localize on egress according to trunk requirements and/or user expectations, using the unmodified caller ID as presented from the PSTN is not compatible with this approach.

Using Partial Caller ID Matching

With this approach, remote destinations are configured as they would be dialed from the system to the PSTN. For example, if the number for the remote destination is 14085557890 and PSTN access from the system requires a 9, then this number should be configured on the Remote Destination configuration page as 914085557890. This approach precludes the need for configuration of a digit prefixing mechanism on the system, but it requires setting the Matching Caller ID with Remote Destination service parameter to Partial Match and setting the Number of Digits for Caller ID Partial Match to the appropriate number of consecutive digits that should be matched against the remote destination caller ID. For example, if the caller ID for a remote destination is 14085557890 and the remote destination is configured as 914085557890, then the Number of Digits for Caller ID Partial Match would ideally be set to 10 or 11. In this example, this parameter could be set to a lower number of digits; however, always ensure that enough consecutive digits are matched so that all configured remote destinations in the

system are matched uniquely. If there is no exact match or if more than one configured remote destination number is matched when using partial caller ID matching, the system treats this as if there is no matching remote destination number, thus requiring the user to enter their remote destination number/ID manually in the case of Mobile Voice Access before providing their PIN. With Enterprise Feature Access, there is no mechanism for the user to enter their remote destination number; therefore, when using this functionality, ensure that only unique matches occur.

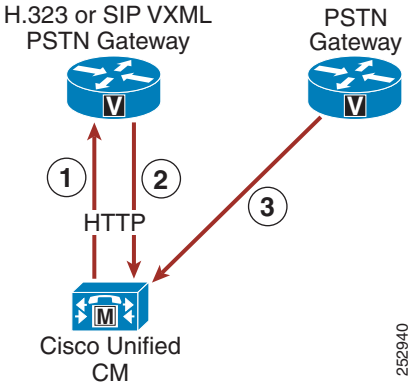
Note If the PSTN service provider sends variable-length caller IDs, using partial caller ID matching is not recommended because ensuring a unique caller ID match for each inbound call might not be possible. In these scenarios, using complete caller ID matching and/or a +E.164 dial plan is the preferred method.

Mobile Voice Access and Enterprise Feature Access Architecture

The architecture of the Mobile Voice Access and Enterprise Feature Access feature is as important to understand as their functionality. Figure 21-26 depicts the message flows and architecture required for Mobile Voice Access and Enterprise Feature Access. The following sequence of interactions and events can occur between Unified CM, the PSTN gateway, and the H.323 or SIP VXML gateway:

1. Unified CM forwards IVR prompts and instructions to the H.323 or SIP VXML gateway via HTTP (see step 1 in Figure 21-26). This provides the VXML gateway with the ability to play these prompts for the inbound Mobile Voice Access callers.
2. The H.323 or SIP VXML gateway uses HTTP to forward Mobile Voice Access user input back to Unified CM (see step 2 in Figure 21-26).
3. The PSTN gateway forwards DTMF digits in response to user or Smart Phone key sequences from the remote destination phone for Enterprise Feature Access two-stage dialing and mid-call features (see step 3 in Figure 21-26).

Figure 21-26 Mobile Voice Access and Enterprise Feature Access Architecture



Note While Figure 21-26 depicts the H.323 or SIP VoiceXML gateway as a separate box from the PSTN gateway, this is not an architectural requirement. Both VoiceXML functionality and PSTN gateway functionality can be handled by the same box, provided there are no requirements for the PSTN gateway to run a protocol other than H.323 or SIP. An H.323 or SIP gateway is required for Mobile Voice Access VoiceXML functionality.

Note Because Cisco IOS XE does not provide native VoiceXML support, the Cisco 4000 Series ISR cannot be used as the VoiceXML gateway for Mobile Voice Access. If the PSTN gateway is a Cisco 4000 Series ISR, you most offload the VoiceXML functionality to a Cisco IOS gateway with native VoiceXML support.

High Availability for Mobile Voice Access and Enterprise Feature Access

The Mobile Voice Access and Enterprise Feature Access features rely on the same components and redundancy mechanisms as the Single Number Reach feature (see High Availability for Single Number Reach, page 21-58). Unified CM Groups are necessary for PSTN gateway registration redundancy. Likewise, PSTN physical gateway and gateway connectivity redundancy should be provided. Redundant access between the PSTN and the enterprise is required for remote destination phones to access Mobile Voice Access and Enterprise Feature Access features in the event of a gateway failure. However, while physical redundancy can and should be provided for the H.323 or SIP VoiceXML gateway, there is no redundancy mechanism for the Cisco Unified Mobile Voice Access service on Unified CM. This service can be enabled and run on the publisher node only. Therefore, if the publisher node fails, Mobile Voice Access functionality will be unavailable. Enterprise Feature Access and two-stage dialing functionality have no such dependency on the publisher and can therefore provide equivalent functionality to mobility users (without the IVR prompts).

Designing Cisco Unified Mobility Deployments

The Cisco Unified Mobility solution delivers mobility functionality via Cisco Unified CM. Functionality includes Single Number Reach, Mobile Voice Access, and Enterprise Feature Access. When deploying this functionality it is important to understand dial plan implications, guidelines and restrictions, and performance and capacity considerations.

Dial Plan Considerations for Cisco Unified Mobility

In order to configure and provision Unified Mobility appropriately, it is important to understand the call routing behavior and dial plan implications of the remote destination profile configuration.

Remote Destination Profile Configuration

When configuring Unified Mobility, you must consider the following two settings on the Remote Destination Profile configuration page:

- Calling Search Space

  This setting combines with the directory number or line-level calling search space (CSS) to determine which partitions can be accessed for mobility dialed calls. This affects calls made by the mobility user from the remote destination phone, including Mobile Voice Access and Enterprise Feature Access two-stage dialing as well as calls made in conjunction with mid-call transfer and conferencing features. Ensure that this CSS, in combination with the line-level CSS, contains all partitions that need to be accessed for enterprise calls originating from a user's remote destination phone. In a +E.164 dial plan using the line-only traditional approach with local route groups, this CSS is not required and can be set to **<None>**.

- Rerouting Calling Search Space

  This setting determines which partitions are accessed when calls are sent to a user's remote destination phone. This applies to all Single Number Reach calls. When a call to a user's enterprise directory number is also sent via Single Number Reach to a user's remote destination, this CSS determines how the system reaches the remote destination phone. For this reason, the CSS should provide access to partitions with appropriate route patterns and gateways for reaching the PSTN or mobile voice network.

When configuring the Remote Destination Profile Rerouting CSS, Cisco recommends that the route patterns within this CSS point to a gateway that is in the same call admission control location as the gateway used to route the inbound call to the user's desk phone. This ensures that a call admission control denial due to insufficient bandwidth between two locations will not occur when routing calls out to the remote destination. Further, because subsequent call admission control checks after the initial Single Number Reach call is routed will not result in a denial if there is insufficient WAN bandwidth, routing the inbound and outbound call legs out a gateway or gateways in the same call admission control location ensures that subsequent desk phone or remote destination pickup operations during this call will not require call admission control, which could result in WAN bandwidth oversubscription.

When using route patterns pointing to route lists that use Standard Local Route Group, the local route group configured on the caller's device pool will be used. In this case the egress gateway for the call leg to the remote destination will be local to the original calling device. For calls coming in from the PSTN, this will help to fulfill the above requirement to use egress gateways in the same call admission control location as the original caller (in this case the incoming gateway).

Likewise, it is equally important to ensure that call admission control denials are minimized when placing two-stage dialed calls. Call admission control denials for two-stage dialed calls can be minimized or avoided by using local route group constructs so that the egress gateway used to route the outbound call leg is chosen by the ingress gateway of the inbound call leg. With this method, the ingress and egress gateways used will be in the same call admission control location. Alternatively, the route patterns within the Remote Destination Profile device-level CCS should point to an egress gateway that is in the same call admission control location as the ingress gateway that handled the inbound call leg to the Mobile Voice Access or Enterprise Feature Access system access number. However, be aware that a subsequent desk phone pickup can result in WAN bandwidth oversubscription if the desk phone is in a different call admission control location than the gateway through which the Mobile Voice Access or Enterprise Feature Access system access numbers are reached.

Automatic Caller ID Matching and Enterprise Call Anchoring

Another aspect of the Unified Mobility dial plan that is important to understand is the system behavior with regard to automatic caller ID identification for inbound calls from configured remote destination phones. Whenever an inbound call comes into the system, the presented caller ID for that call is compared against all configured remote destination phones. If a match is found, the call will automatically be anchored in the enterprise, thus allowing the user to invoke mid-call features and to pick up in-progress calls at their desk phone. This behavior occurs for all inbound calls from any mobility user's remote destination phone, even if the inbound call is not originated as a mobility call using Mobile Voice Access or Enterprise Feature Access.

Note Automatic inbound caller ID matching for configured remote destination numbers is affected by whether the Matching Caller ID with Remote Destination service parameter is set to Partial or Complete Match. See Remote Destination Configuration and Caller ID Matching, page 21-65, for more information about this setting.

In addition to automatic enterprise call anchoring, inbound and outbound call routing must also be considered when a configured remote destination phone is calling into the enterprise. Inbound call routing for calls from configured remote destinations occurs in one of two ways, depending on the setting of the service parameter Inbound Calling Search Space for Remote Destination. By default, this service parameter is set to **Trunk or Gateway Inbound Calling Search Space**. With the service parameter set to the default value, inbound calls from configured remote destinations will be routed using the Inbound Calling Search Space (CSS) of the PSTN gateway or trunk on which the call is coming in. If, on the other hand, the parameter Inbound Calling Search Space for Remote Destination is set to the value **Remote Destination Profile + Line Calling Search Space**, inbound calls coming from remote destinations will bypass the Inbound CSS of the PSTN gateway or trunk and will instead be routed using the associated Remote Destination Profile CSS (in combination with the line-level CSS).

Given the nature of inbound call routing from remote destination phones, it is important to make sure that calling search spaces are configured appropriately in order to provide access for these inbound calls to any partitions required for reaching internal enterprise phones, thus ensuring proper call routing from remote destination phones.

Note Incoming calls that do not come from a configured remote destination phone are not affected by the Inbound Calling Search Space for Remote Destination service parameter because they will always use the trunk or gateway inbound CSS.

Outbound call routing for Mobile Voice Access or Enterprise Feature Access calls always uses a concatenation of the Remote Destination Profile line CSS and device-level CSS, therefore it is important to make sure that these calling search spaces are configured appropriately in order to provide access to any route patterns necessary for off-net or PSTN access, thus ensuring proper outbound call routing from remote destination phones.

Intelligent Session Control and Ring All Shared Lines

The Intelligent Session Control feature enables automatic call anchoring for enterprise-originated calls made directly to configured remote destination numbers. Normally, mobility call anchoring is dependent exclusively on calls made to or on behalf of a user's enterprise number. The system already anchors externally originated calls made by enterprise two-stage dialing because these call are routed as internal calls. With the Intelligent Session Control feature enabled, the system will also anchor internally originated calls made directly to configured remote destinations.

This feature is enabled by setting the Reroute Remote Destination Calls to Enterprise Number service parameter to True. By default, this service parameter is set to False and the feature is disabled. When the feature is enabled, not only will the system route the call to the dialed remote destination by way of the PSTN, but it will also automatically anchor the call inside the enterprise gateway. By anchoring these types of calls, the system enables the called mobile user to invoke mid-call features and desk phone pickup or session handoff.

As an example, assume that the Intelligent Session Control feature has been enabled and that a mobility-enabled user has a remote destination number configured as 408 555 1234, which corresponds to their mobile number. If another system user dials the mobility-enabled user's remote destination number (408 555 1234) from their desk phone, the system will route the call through the PSTN to the remote destination and will simultaneously anchor the call in the enterprise gateway. Once the call is set up and anchored, the called mobility-enabled user now has the ability to invoke mid-call features such as hold, transfer, and conference, as well as the ability to perform a desk phone pickup or session handoff.

Taking this same example and assuming instead that the Intelligent Session Control feature is disabled, then when a system user dials the mobility-enabled user's remote destination directly from a desk phone inside the enterprise, the call will still be routed to the called remote destination through the PSTN; however, the call will not be anchored. As a result, the mobile user would not be able to invoke mid-call hold or transfer and would have no ability to perform a desk phone pickup or session handoff.

When enabling this feature, it is important to understand the implications to dial plan configuration and call routing. To invoke the feature, the number dialed by an internal user to reach a remote destination number on the PSTN (including any required PSTN steering digits) must match the remote destination (or mobility identity) number as it is configured on the system. For example, if the remote destination number is configured on the system as 408 555 1234 but internal users must normally dial PSTN steering digits 91 in addition to the number they are calling, then rerouting and resulting enterprise call anchoring will not occur. This is because the user dialed 91 408 555 1234 to reach the remote destination on the PSTN but the remote destination was configured as 408 555 1234, so there is no match.

For this feature to function properly, matching must occur between the configured remote destination and the number that must be dialed to reach this remote destination on the PSTN. To ensure that this matching happens, set the service parameter Matching Caller ID with Remote Destination to **Partial Match**. By setting this parameter to Partial Match and then specifying the number of digits to partially match using the Number of Digits for Caller ID Partial Match service parameter, it is still possible to match the configured remote destination number with the dialed number even if it contains PSTN steering digits.

Using the previous example and assuming that system has been set to use partial match on ten digits, the dialed number 9 1 408 555 1234 can be matched to the configured remote destination 408 555 1234. This is because, with partial matching, the system attempts to match the same number of digits as specified by the Number of Digits for Caller ID Partial Match, which in this case is ten digits. The system attempts to match the two numbers by matching digits from right to left. The last ten digits of the dialed number 9 1 408 555 1234 are 408 555 1234, and these ten digits match the ten digits of the configured remote destination (408 555 1234). In this example, the resulting call is anchored in the enterprise and the called mobile user is able to invoke mid-call features and perform desk phone pickup or session handoff.

At first glance it might appear that an easier way to handle this feature would be to configure remote destination or mobility identity numbers that include any required PSTN steering digits. However, when configuring these numbers with required PSTN steering digits, if you do not also configure partial caller ID matching, the system will not be able to perform automatic caller ID matching and enterprise anchoring for inbound calls from configured remote destinations or mobility identities. In the previous example, if the remote destination number had been configured as 9 1 408 555 1234 and complete caller ID matching had been used, an inbound call from the remote destination would present caller ID of 408 555 1234 and a match would not occur, meaning the inbound call from the remote destination would not be anchored as expected.

Based on this potential for mismatch between dialed numbers for outbound calls and configured remote destination numbers for inbound calls, Cisco recommends enabling partial (rather than complete) caller ID matching when using the Intelligent Session Control feature for all deployments that require one or more steering digits to reach the PSTN. This ensures that calls made directly to the remote destination number using PSTN steering digits are still matched and anchored. On the other hand, if steering digits are not required to reach the PSTN and users are able to dial the full E.164 number to route calls to the PSTN, then Cisco recommends the complete caller ID matching setting because the remote destination is configured to match the caller ID and is the same number as dialed by internal users to reach the remote destination or mobility identity on the PSTN.

When enabling the Intelligent Session Control feature, it is also important to understand the behavior of the enterprise and remote destination lines during the reroute feature operation. On call reroute, remote destination line settings Do Not Disturb (DND), Access Lists and Time of Day call filtering, and the

Delay Before Ringing Timer are ignored. All reroute calls are routed unfiltered and immediately. Enterprise desk phone line settings are also ignored or bypassed by default. However, Call Forward All settings on the enterprise desk phone line can be honored during reroute feature operation by setting the Ignore Call Forward All on Enterprise DN service parameter to False. If this parameter is set to False, on reroute operation, calls will not be routed to the remote destination if the enterprise desk phone line has a call-forward-all destination set. Instead, the call will be routed to the call-forward-all destination. By default, this service parameter is set to True, and call-forward-all settings on enterprise desk phone lines are ignored.

Intelligent Session Control functionality may be further enhanced by using the Ring All Shared Lines feature. This feature is enabled by setting the Ring All Shared Lines service parameter to True. By default, this service parameter is set to True and the feature is enabled. However, the Ring All Shared Lines feature is dependent on the Intelligent Session Control feature, which must also be enabled in order use the Ring All Shared Line functionality. When the Ring All Shared Lines and Intelligent Session Control features are both enabled, not only will the system route internally originated calls to the dialed remote destination by way of the PSTN, but all of the user's other shared-line devices will also receive the call. This includes the user's enterprise desk phone as well as other configured remote destinations. The called user will then be able to answer the incoming call on any of their devices and the call will be anchored in the enterprise.

Note If Ring All Shared Lines is enabled, mobile client devices will not receive calls at the cellular voice interface of the device when the device is registered to Unified CM.

Caller ID Transformations

Any calls made into the cluster by configured remote destination numbers will automatically have their caller ID or calling number changed from the calling remote destination phone number to the enterprise directory number of the associated desk phone. For example, if a remote destination phone with number 408 555-7890 has been configured and associated to a user's enterprise desk phone with number 555-1234, then any call from the user's remote destination phone destined for any directory number in the cluster will automatically have the caller ID changed from the remote destination number of 408 555-7890 to the enterprise directory number of 555-1234. This ensures that the active call caller ID display and call history log caller ID reflect a mobility user's enterprise desk phone number rather than their mobile phone number, and it ensures that any return calls are made to the user's enterprise number, thus anchoring those calls within the enterprise.

Likewise, calls from a remote destination phone to external PSTN destinations and anchored in the enterprise via Mobile Voice Access or Enterprise Feature Access two-stage dialing, or those calls forked to the PSTN as a result of Single Number Reach, will also have caller ID changed from the calling remote destination phone number to the associated enterprise directory number.

Finally, in order to deliver the calling party number as an enterprise DID number rather than an enterprise directory number to external PSTN phones, calling party transformation patterns can be used. By using calling party transformation patterns to transform caller IDs from enterprise directory numbers to enterprise DIDs, return calls from external destinations will be anchored within the enterprise because they will be dialed using the full enterprise DID number. For more information about these transformations and dial plan implications, see Special Considerations for Cisco Unified Mobility, page 14-86.

Intelligent Proximity for Mobile Voice and Unified Mobility Interactions

The Intelligent Proximity for Mobile Voice feature on the Cisco DX Series endpoints and Cisco IP Phones 8851 and 8861 is compatible with the Unified Mobility feature set, including single number reach (SNR), remote destination and desk phone pickup, two-stage enterprise dialing, and mobile voicemail avoidance. For more information on Intelligent Proximity for Mobile Voice and Bluetooth pairing on the DX Series endpoints and 8851 and 8861 IP Phones, see Intelligent Proximity, page 8-13.

Guidelines and Restrictions for Unified Mobility

Note The Cisco Unified Mobility solution is verified with only Cisco equipment. This solution may also work with other third-party PSTN gateways and Session Border Controllers (SBCs), but each Cisco Mobility feature is not guaranteed to work as expected. If you are using this solution with third-party PSTN gateways or SBCs, Cisco technical support may not be able to resolve problems that you encounter.

The following guidelines and restrictions apply with regard to deployment and operation of Single Number Reach within the Unified CM telephony environment:

- Single Number Reach is supported only with PRI TDM PSTN connections. T1 or E1-CAS, FXO, FXS, and BRI PSTN connections are not supported. This PRI requirement is based on the fact that Cisco Unified CM must receive expeditious answer and disconnect indication from the PSTN in order to ensure full feature support. Answer indication is needed in order for Cisco Unified CM to stop ringing the desk phone and other remote destinations when a Single Number Reach call is answered at a particular remote destination. In addition, answer indication is required in order to support the single enterprise voicemail box feature. Finally, disconnect indication is required for desk phone pickup. A PRI PSTN connection will always provide answer or disconnect indication.
- Single Number Reach is also supported over SIP trunk VoIP PSTN connections. Use of Cisco IOS Unified Border Element is recommended as the demarcation point between the Unified CM SIP trunk and the service provider trunk. A VoIP-based PSTN connection is still able to provide expeditious answer and disconnect indication to Unified CM due to the end-to-end signaling path provided by VoIP-based PSTN connections.
- Single Number Reach can support up to two simultaneous calls per user. Any additional calls that come in are automatically transferred to the user's voicemail.
- Single Number Reach does not work with Multilevel Precedence and Preemption (MLPP). If a call is preempted with MLPP, Single Number Reach features are disabled for that call.
- Single Number Reach services do not extend to video calls. A video call received at the desktop phone cannot be picked up on the cellular phone.
- Remote destinations must be Time Division Multiplex (TDM) devices or off-system IP phones on other clusters or systems. You cannot configure IP phones within the same Unified CM cluster as remote destinations.
- Mobile Voice Access VoiceXML capabilities are not supported with Cisco IOS XE software. Because there is no native VoiceXML support with Cisco IOS XE, the Cisco 4000 Series ISR cannot serve as a VXML gateway for Mobile Voice Access. Instead, deploy a separate H.323 Cisco IOS gateway for VoiceXML capabilities and configure Mobile Voice Access for hairpinning.

For additional guidelines and restrictions, refer to the information on Cisco Unified Mobility in the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Capacity Planning for Cisco Unified Mobility

Cisco Unified Mobility supports a maximum of 40,000 remote destinations or mobility identities per Unified CM cluster. The maximum number of mobility-enabled users would thus be 40,000 users, assuming a single remote destination or mobility identity per user. As the number of remote destinations or mobility identities per user increases, the number of supported mobility-enabled users decreases.

Note A mobility-enabled user is defined as a user that has a remote destination profile and at least one remote destination or a mobile client device and a mobility identity configured.

Note A mobility identity is configured just like a remote destination within the system, and it has the same capacity implications as a remote destination. Unlike a remote destination, however, the mobility identity is associated directly to a phone device rather than a remote destination profile. The mobility identity applies only to dual-mode mobile client devices running Cisco Jabber.

Scalability and performance of Cisco Unified Mobility ultimately depends on the number of mobility users, the number of remote destinations or mobility identities each user has, and the busy hour call attempt (BHCA) rates of those users. Multiple remote destinations per user and/or high BHCA per user can result in lower capacity for Cisco Unified Mobility. For more information on Cisco Unified Mobility sizing, including Unified CM server node capacities and hardware specific per-node and per-cluster capacities, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Design Considerations for Cisco Unified Mobility

Observe the following design recommendations when deploying Unified Mobility:

- Ensure that the PSTN gateway protocol is capable of out-of-band DTMF relay or allocate media termination points (MTPs) in order to covert in-band DTMF to out-of-band DTMF. When using Cisco IOS gateways for PSTN connectivity, out-of-band DTMF relay will be supported. However, third-party gateways might not support a common out-of-band DTMF method, and as a result an MTP might be required. In order to use Enterprise Feature Access Two-Stage Dialing and mid-call features, DTMF digits must be received out-of-band by Cisco Unified CM.

  Note When relying on MTP for converting in-band DTMF to out-of-band DTMF, be sure to provide sufficient MTP capacity. If heavy or frequent use of Enterprise Feature Access Two-Stage Dialing or mid-call features is anticipated, Cisco recommends a hardware-based MTP or Cisco IOS software-based MTP.

- Prior to deploying Unified Mobility, it is important to work with the PSTN provider to ensure the following:
  - Caller ID is provided by the service provider for all inbound calls to the enterprise. This is a requirement if Enterprise Feature Access Two-Stage Dialing or mid-call transfer, conference, and directed call park features are needed.
  - Outbound caller ID is not restricted by the service provider. This is a requirement if there is an expectation that mobility-enabled users will receive the caller ID of the original caller at their remote destination rather than a general enterprise system number or other non-meaningful caller ID.

Note Some providers restrict outbound caller ID on a trunk to only those DIDs handled by that trunk. For this reason, a second PRI trunk that does not restrict caller ID might have to be acquired from the provider. To obtain an unrestricted PRI trunk, some providers might require a signed agreement from the customer indicating they will not send or make calls to emergency numbers over this trunk.

Note Some providers allow unrestricted outbound caller ID on a trunk as long as the Redirected Dialed Number Identification Service (RDNIS) field or SIP Diversion Header contains a DID handled by the trunk. The RDNIS or SIP Diversion Header for forked calls to remote destinations can be populated with the enterprise number of the user by checking the Redirecting Number IE Delivery - Outbound check box on the gateway or trunk configuration page. Contact your service provider to determine if they honor the RDNIS or SIP Diversion Header and allow unrestricted outbound caller ID.

- Because mobility call flows typically involve multiple PSTN call legs, planning and allocation of PSTN gateway resources is extremely important for Unified Mobility. In cases where there are large numbers of mobility-enabled users, PSTN gateway resources will have to be increased. The following methods are recommend to minimize or reduce PSTN utilization:
  - Limit the number of remote destinations per mobility-enabled user to one (1). This will reduce the number of DS0s that are needed to extend the inbound call to the user's remote destination. One DS0 is consumed for each configured remote destination when a call comes into the user's enterprise directory number, even if the call is not answered at one of the remote destinations. Note that a DS0 per remote destination may be used for as long as 10 seconds, even if the call is not answered at the remote destination.
  - Use access lists to block or restrict the extension of calls to a particular remote destination based on incoming caller ID. Because access lists can be invoked based on the time of day, this eliminates the need for repeated updates of access lists by the end-user or the administrator.
  - Educate end-users to disable Single Number Reach when not needed, to further eliminate DS0 utilization when a call comes in for that user's enterprise number. If Single Number Reach is disabled, incoming calls will still ring the desk phone and will still forward to enterprise voicemail if the call goes unanswered.
- Due to the potential for call admission control denials resulting from insufficient WAN bandwidth between locations and the possibility that a desk phone pickup or remote destination pickup might result in WAN bandwidth over-subscription, Cisco recommends configuring Remote Destination Profile CSS and Rerouting CSS so that route patterns within these CSSs point to gateways that are located within the same call admission control location as the gateway on which the inbound call leg comes in. For more information, see Remote Destination Profile Configuration, page 21-68.
- If you enable the Intelligent Session Control feature in deployments where PSTN steering digits must be dialed to access the PSTN, Cisco recommends setting the Matching Caller ID with Remote Destination service parameter to **Partial Match** and configuring the appropriate number of digits (Number of Digits for Caller ID Partial Match service parameter) to achieve a partial match of configured remote destinations or mobility identities. This will ensure proper functioning of the Intelligent Session Control feature and the mobility automatic caller ID matching and anchoring features.

Cisco Mobile Clients and Devices

As the prevalence of mobile users, mobile phones, and mobile carrier services continues to increase, the ability to use a single device for voice, video, and data services both inside and outside the enterprise becomes increasingly attractive. Mobile devices, including dual-mode smartphones and the clients that run on them, afford an enterprise the ability to provide customized voice, video, and data services to users while inside the enterprise and to leverage the mobile carrier network as an alternate connection method for general voice and data services. By enabling voice, video, and data services inside the enterprise and providing network connectivity for mobile client devices, enterprises are able to provide these services locally or remotely at reduced connectivity costs. For example, voice over IP (VoIP) calls made on the enterprise network will typically incur less cost than those same calls made over the mobile voice network.

In addition to providing voice and video over IP (VVoIP) capabilities, these mobile clients and devices enable mobile users to access and leverage other back-end collaboration applications and services. Other services and applications that can be leveraged through Cisco mobile clients and services include enterprise directory, enterprise voicemail, and XMPP-based enterprise IM (instant messaging) and presence. Further, these clients and devices can be deployed in conjunction with Cisco Unified Mobility so that users can leverage additional features and functions with their mobile device, such as Single Number Reach, enterprise two-stage dialing through Mobile Voice Access or Enterprise Feature Access, and single enterprise voicemail box.

This section examines mobile client architecture and common functions and features provided by Cisco mobile clients and devices, including remote secure attachment and handoff considerations related to moving an active voice call between the enterprise WLAN network and the mobile voice network. After covering the general mobile client solution architecture and features and functions, this section provides coverage of various capabilities and integration considerations for the following specific mobile clients and devices:

- Cisco Jabber — A mobile client available for Android and Apple iOS mobile devices, including iPhone and iPad, providing the ability to make voice and/or video calls over IP on the enterprise WLAN network or over the mobile data network as well as the ability to access the corporate directory and enterprise voicemail services and XMPP-based enterprise IM and presence.
- Cisco Spark — A mobile client available for Android and Apple iOS devices, including iPhone and iPad, providing 1-to-1 and 1-to-many cloud-based collaboration rooms enabling voice and/or video calls over IP, secure persistent messaging, and file sharing.
- Cisco WebEx Meetings — A mobile client available for Android, BlackBerry, Windows Mobile, and Apple iOS devices including iPhone and iPad, enabling users to attend and participate in Cisco WebEx meetings while mobile.
- Cisco AnyConnect Mobile — A mobile client available for Android and Apple iOS devices, enabling secure remote VPN connectivity to the enterprise for access to on-premises collaboration applications and services even when the user is outside of the enterprise.

In addition, this section discusses high availability and capacity planning considerations for Cisco mobile clients and devices.

Cisco Mobile Clients and Devices Architecture

Cisco mobile clients are deployed on a wide range of mobile devices including tablets and handheld devices with only IP-based network connectivity capabilities (IEEE 802.11 wireless local area network or mobile provider data network) and dual-mode phones, which contain two physical interfaces or radios that enable the device to connect to both mobile voice and data carrier networks by means of traditional cellular or mobile network technologies and to connect to wireless local area networks (WLANs) using 802.11. Cisco mobile clients and devices enable on-premises data and real-time traffic (voice and video) connectivity through an 802.11 WLAN. In addition, these clients and devices provide remote data and real-time traffic (voice and video) connectivity to the enterprise through public or private WLANs or over the mobile data network. For devices with provider cellular voice radios, voice connectivity may also be enabled through the mobile voice network and PSTN.

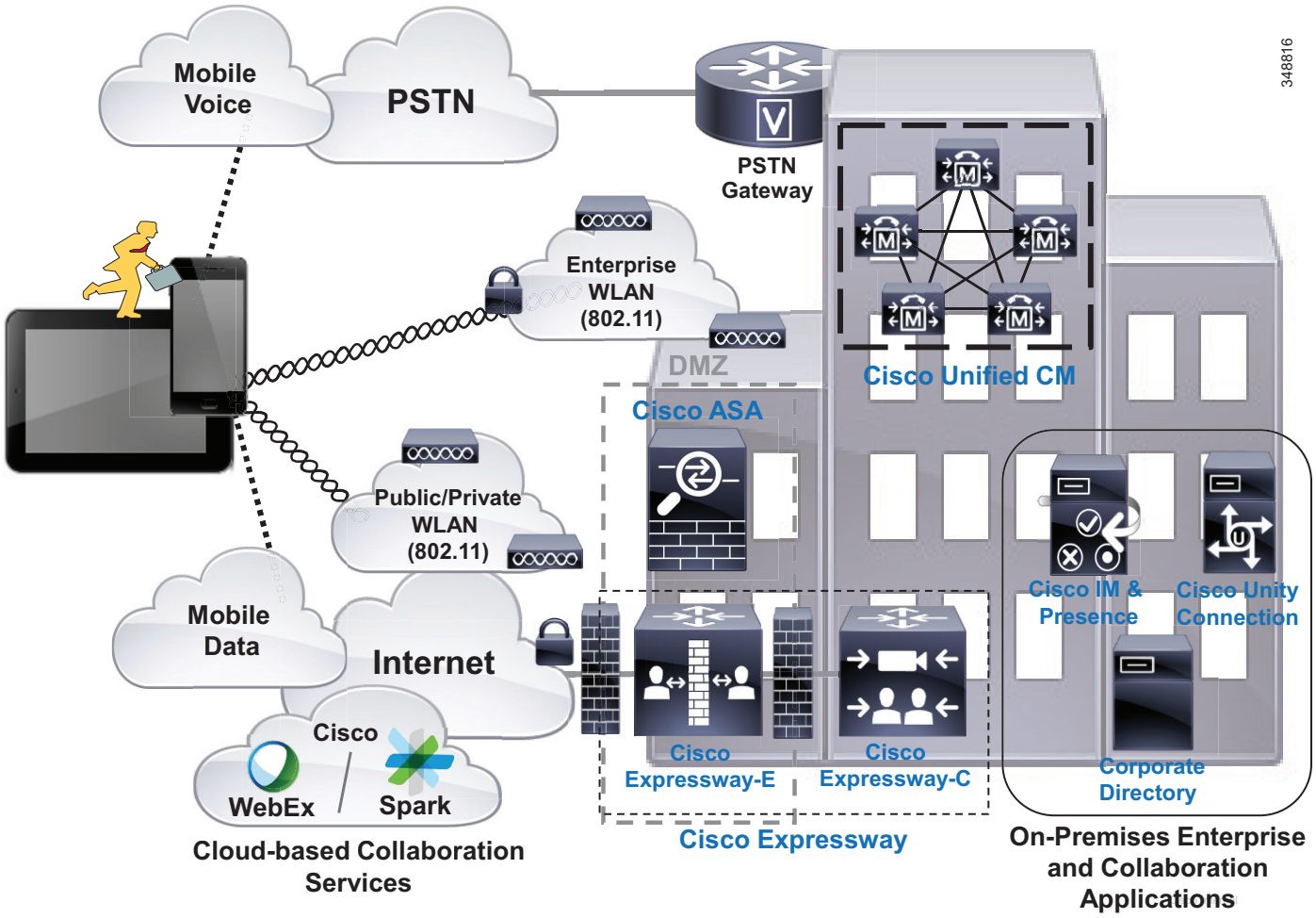
Note The use of the term *dual-mode phone* in this section refers specifically to devices with 802.11 radios in addition to the cellular radio for carrier voice and data network connectivity. Dual-mode devices that provide Digital Enhanced Cordless Telecommunications (DECT) or other wireless radios and/or multiple cellular radios are outside the scope of this section.

Figure 21-27 depicts the basic Cisco mobile clients and devices solution architecture for connecting and enabling mobile client devices for Cisco Collaboration deployments. For voice and video services, mobile client devices associate to the enterprise WLAN or connect over the Internet (from a public or private WLAN hot spot or the mobile data network), and the Cisco mobile client registers to Cisco Unified CM as an enterprise phone using the Session Initiation Protocol (SIP). Once registered, the client device relies on the underlying enterprise Cisco IP telephony network for making and receiving calls. When the mobile device is connected to the enterprise network and the client is registered to Unified CM, the device is reachable through the user's enterprise number. Any inbound calls to the user's enterprise number will ring the mobile client device. If the user has a Cisco IP desk phone, then the mobile client registration enables a shared line instance for the user's enterprise number so that an incoming call rings both the user's desk phone and the mobile device. When unregistered, the mobile client device will not receive incoming enterprise calls unless the mobile device has an active cellular voice radio, the user has been enabled for Cisco Unified Mobility, and Single Number Reach has been turned on for the user's mobile phone number. In these scenarios the mobile voice network and PSTN are used for making and receiving voice-only calls.

Unified Mobility features such as Single Number Reach are not compatible with tablets and other mobile client devices that do not have cellular voice radios because these non-dual-mode devices do not have a native PSTN reachable number. Non-dual-mode devices are able to make and receive enterprise calls only when connected to the enterprise and registered to the enterprise call control system.

As shown in Figure 21-27, when attached to the enterprise, Cisco mobile clients and devices can also communicate directly with other back-end application servers such as the corporate directory, Cisco Unity Connection enterprise voicemail system, and the Cisco IM and Presence Service for access to additional enterprise collaboration services such as messaging and presence. Cisco mobile clients and devices also integrate with cloud-based collaboration services such as Cisco WebEx, which delivers IM and presence and web conferencing services.

Figure 21-27 Cisco Mobile Clients and Devices Architecture



Note The voice and video quality of calls will vary depending on the Wi-Fi or mobile data network connection. Cisco Technical Assistance Center (TAC) is not able to troubleshoot connectivity or voice and video quality issues over 3G/4G mobile data networks or non-corporate Wi-Fi networks.

Dual-mode mobile client devices must be capable of dual transfer mode (DTM) in order to be connected simultaneously to both the mobile voice and data network and the WLAN network. This allows the device to be reachable and able to make and receive calls on both the cellular radio and WLAN interface of the device. In some cases proper mobile client operation might not be possible if mobile voice and data networks do not support dual-connected devices.

Voice and Video over Wireless LAN Network Infrastructure

Before considering the various mobile client device features and functions and the impact these features and functions have on the enterprise telephony infrastructure, it is critical to plan and deploy a finely tuned, QoS-enabled, and highly available WLAN network. Because dual-mode phones and other mobile

devices rely on the underlying WLAN infrastructure for carrying both critical signaling and other traffic for setting up calls and accessing various applications as well as the real-time voice and video media traffic, deploying a WLAN network optimized for both data and real-time media traffic is necessary. A poorly deployed WLAN network will be subjected to large amounts of interference and diminished capacity, leading not only to poor voice and video quality but in some cases dropped or missed calls. This will in turn render the WLAN deployment unusable for making and receiving calls. Therefore, when deploying dual-mode phones and other mobile devices, it is imperative to conduct a WLAN radio frequency (RF) site survey before, during, and after the deployment to determine appropriate cell boundaries, configuration and feature settings, capacity, and redundancy to ensure a successful deployment of voice and video over WLAN. Each mobile device type and/or client should be tested on the WLAN deployment to ensure proper integration and operation prior to a production deployment. Using a WLAN that has been deployed and configured to provide optimized real-time traffic over WLAN services (such as the Cisco Unified Wireless Network), including quality of service, will ensure a successful mobile client device deployment.

Cisco recommends relying on the 5 GHz WLAN band (802.11a/n/ac) whenever possible for connecting mobile clients and devices capable of generating voice and video traffic. 5 GHz WLANs provide better throughput and less interference for voice and video calls.

For more information on voice and video over WLAN deployments and wireless device roaming, see Wireless Device Roaming, page 21-6.

Note While dual-mode phones and other mobile client devices are capable of connecting back to the enterprise through the Internet for call control and other Unified Communications services, Cisco cannot guarantee voice and video quality or troubleshoot connectivity or voice and video quality issues for these types of connections. These types of connections include remote connections to the enterprise through public or private WLAN access points (APs) or hot spots or through the mobile data network. Cisco recommends an enterprise class voice and video-optimized WLAN network for connecting dual-mode phones and other mobile client devices. Most public and private WLAN APs and hot spots are tuned for data applications and devices. In these cases, the AP radios are turned to maximum power, and dynamic-power control results in devices enabling maximum power on network attachment, which allows for larger client capacities. While this may be ideal for data applications that are capable of retransmitting dropped or lost packets, for real-time traffic applications this can result in poor voice and video quality due to the potential for large numbers of dropped packets. Likewise, mobile provider data networks are susceptible to congestion and/or dropped connections, which can also result in poor call quality and dropped calls.

Cloud or Off-Premises Collaboration Infrastructure

Cisco WebEx and Cisco Spark, cloud services available from the Cisco, do not require any hardware to be deployed on the enterprise premises. All services (audio, video, messaging, file and content sharing, and meeting and collaboration room information) are securely hosted in the Internet or the cloud. This means that all the content, voice, and video traffic from every client traverses the internet and is mixed and managed in the Cisco Collaboration Cloud.

The Cisco Collaboration Cloud infrastructure provides WebEx and Cisco Spark capabilities to mobile clients and devices, including:

- WebEx Meetings, which provides web-enabled voice and video conferencing with content sharing.
- WebEx Messenger, which provides XMPP IM and presence as well as point-to-point audio and video calling.
- Cisco Spark, which provides 1-to-1 and 1-to-many collaboration rooms with voice and video calling, messaging, and file sharing.

Mobile Client and Device Quality of Service

Cisco mobile client applications and devices generally mark Layer 3 QoS packet values in accordance with Cisco collaboration QoS marking recommendations. Table 21-3 summarizes these markings.

Table 21-3 Cisco Mobile Client Layer 3 QoS Markings

Traffic Type	Layer 3 Marking	
	DSCP[1]	PHB[2]
Voice media (audio only)	DSCP 46	PHB EF
Video media (audio and video)	DSCP 34	PHB AF41
Call Signaling	DSCP 24	PHB CS3

1. Differentiated Services Code Point
2. Per-Hop Behavior

Cisco mobile client Layer 2 802.11 WLAN packet marking (User Priority, or UP) presents challenges given the various mobile platform and firmware restrictions. Because Cisco mobile clients run on a variety of mobile devices, Layer 2 wireless QoS marking is inconsistent and therefore Layer 2 wireless QoS marking cannot be relied on to provide appropriate treatment to traffic on the WLAN.

Despite appropriate mobile client application Layer 3 or even Layer 2 packet marking, mobile devices present many of the same challenges as desktop PCs in terms of generating many different types of traffic, including both data and real-time traffic. Given this, mobile devices generally fall into the untrusted category of collaboration endpoints. For deployments where mobile client devices are not considered trusted endpoints, packet marking or re-marking based on traffic type and port numbers is required to ensure that network priority queuing and dedicated bandwidth is applied to appropriate traffic. In addition to re-marking the mobile device traffic, Cisco recommends using network-based policing and rate limiting to ensure that the mobile client devices do not consume too much network bandwidth.

Alternatively, given appropriate Cisco mobile client Layer 3 marking and assuming mobile client devices are trusted, Cisco mobile client traffic will be queued appropriately as it traverses the enterprise network by using priority voice queuing and dedicated video media and call signaling bandwidth queues.

Cisco Mobile Clients and Devices Features and Functions

Cisco mobile clients and devices provide a range of features and functions. While features and operations may vary from device to device, the common operations and behaviors described in this section apply to all non-cloud-based Cisco mobile clients.

Enterprise Call Routing

Because Cisco mobile clients and devices are capable of making and receiving calls using the enterprise telephony infrastructure and call control services, it is important to understand the nature and behavior of call routing as it pertains to mobile client devices.

Inbound Call Routing

When mobile clients and devices register to Unified CM as an enterprise device with enterprise number, the mobile device rings when incoming calls to the system are destined for the user's enterprise number. This occurs for incoming calls originated on the PSTN or from other Unified CM clusters or enterprise IP telephony systems as well as for incoming calls originated within the Unified CM cluster by other

users. If the mobile client device user has other devices or clients that are also associated to the enterprise number, these devices will also ring as shared lines; and once the call is answered at one of the devices or clients, ringing of all other devices and clients ceases.

In scenarios where a user has been enabled for Cisco Unified Mobility, and when Single Number Reach is enabled for the user's dual-mode mobile phone number, the incoming call may also be extended to the mobility identity corresponding to the user's mobile phone number. However, this depends on whether the mobile device is connected to the enterprise WLAN network or attached to the enterprise network through a secure connection and registered to Unified CM. In situations in which the device is connected to the enterprise network directly or through a secure remote connection, an incoming call to the user's enterprise number will not be extended by Single Number Reach to the mobility identity of the mobile device even if Single Number Reach is enabled on for this mobile number. The reason an incoming call to the enterprise number is not extended to the mobility identity of a dual-mode mobile device when it is registered to Unified CM is that the system is aware the device is connected to the enterprise network and available. Thus, in order to reduce utilization of enterprise PSTN resources, Unified CM does not extend the call to the dual-mode mobile phone's mobile voice network interface through the PSTN. Instead, only the WLAN or mobile data network interface corresponding to the enterprise number receives the call.

Note In cases where dial via office is enabled (see Dial Via Office, page 21-86), even if the client is registered, Unified CM will extend inbound calls to the user's mobile number using Single Number Reach rather than via VoIP to the enterprise number.

For situations in which the mobile device is not connected to the enterprise network directly or through a secure remote connection or is not registered to Unified CM, incoming calls to the enterprise number will be extended to the dual-mode mobile phone number per the configured mobility identity, assuming that the user has been enabled for Unified Mobility and that Single Number Reach for the mobility identity is turned on. For more information on integration of mobile clients and devices with Unified Mobility, see Interactions Between Cisco Jabber and Cisco Unified Mobility, page 21-107.

The same behavior and logic described above also applies with the Ring All Shared Lines feature. If this feature is enabled, calls are extended to the mobility identity or cellular number only when the dual-mode mobile client device is *not* registered to Unified CM. For more information on the Ring All Share Line feature, see Intelligent Session Control and Ring All Shared Lines, page 21-70.

In all cases, incoming calls made directly to the dual-mode device's mobile network phone number will always be routed directly to the mobile voice interface of the dual-mode device on the mobile network, unless the provider network or device settings are such that calls are not extended to the device by the mobile network. This is considered appropriate behavior because these calls were not made to the user's enterprise number. These would be considered personal calls, and as such should not be routed through the enterprise.

Note Mobile client devices that do not have cellular voice radios, such as tablet devices, are not dual-mode devices and as such cannot be reached on a mobile voice network interface. These devices can be reached only at the enterprise number by voice-over-IP.

Outbound Call Routing

For outbound calls from the dual-mode mobile device, the interface used depends on the location and connectivity of the device at that particular time. If the dual-mode device is not connected to the enterprise and not registered to Unified CM, then calls are routed by the cellular voice radio interface to the mobile voice network as usual. However, when connected to the enterprise and registered to Unified CM, the mobile device should make all calls through the enterprise telephony infrastructure. If

no enterprise connectivity is available or the mobile client is unregistered, then outbound calling is not possible from the enterprise number, and instead calls would have to use the mobile number of the mobile client device for making calls over the mobile voice network. Alternatively, users may use the two-stage dialing features provided with Cisco Unified Mobility (see Mobile Voice Access and Enterprise Feature Access, page 21-59).

Dial Plan

The enterprise dial plan determines the dialing behavior of the mobile client device when it is connected to the enterprise and registered to Unified CM. For example, if the enterprise dial plan is configured to allow abbreviated dialing to reach internal extensions, then a mobile device registered to Unified CM can use this abbreviated dialing. While it is certainly a convenience for dual-mode mobile phone users to be able to dial within the enterprise using enterprise dialing habits and abbreviated dialing as well as site-based and/or PSTN steering digits for outbound calls, it is also a somewhat unnatural dialing scheme because mobile phone users typically dial numbers for outgoing calls on their mobile phone by using full E.164 dial strings since this is what is expected by the mobile voice network for outbound calling.

The enterprise dialing experience for an end-user is ultimately up to the enterprise policies and administrator of the enterprise telephony deployment. However, for dual-mode mobile devices, Cisco recommends normalizing required dialing strings for dual-mode client devices so that user dialing habits are maintained whether the device is connected to the enterprise network and registered to Unified CM or not. Because dialing on the mobile voice network is typically done using full +E.164 (with a preceding '+') and mobile phone contacts are typically stored with full +E.164 numbers, Cisco recommends configuring the enterprise dial plan to accommodate full +E.164 with preceding '+' for dual-mode mobile devices. When the dial plan is configured within Unified CM to handle this type of outbound dialing for dual-mode phones, it is possible for users to store a single set of contacts on the phone in the +E.164 format and, when dialed from these contacts or manually using the full +E.164 number, calls will always be routed to the appropriate destination, whether the device is connected to the enterprise network directly or over secure remote connection and registered to Unified CM or connected only to the mobile voice network. Configuring the enterprise dial plan in this manner provides the best possible end-user dialing experience so that users' mobile device dialing habits are maintained and they do not have to be aware of whether the device has enterprise connectivity and is registered to Unified CM.

To achieve normalized dialing from dual-mode phones, whether connected to the enterprise or just the mobile voice network, configure the dial plan within Unified CM with the following considerations in mind:

- Ensure that the enterprise dial plan is capable of handling dial strings from dual-mode phones typically used on the mobile voice network. For example, the dial plan should be configured to handle the following strings, which might be dialed from a mobile phone to reach a particular phone through the mobile voice network: +1 408 555 1234, 408 555 1234. Supporting the latter 10-digit dialing method (for example, 408 555 1234) might potentially overlap with other dialing habits such as abbreviated intra-site dialing. In that case the administrator has to decide which of the colliding dialing habits (10-digit dialing or abbreviated intra-site) should be available for dual-mode phones registered to the enterprise network. The set of dialing habits supported on dual-mode phones often differs from the set of dialing habits supported on regular endpoints.
- For calls destined for other enterprise numbers, systems configured for abbreviated dialing should be capable of modifying dial strings and rerouting to enterprise extensions as appropriate. For example, assuming the enterprise dial plan is based on five-digit internal dialing, the system should be configured to handle call routing to an enterprise extension so that a call to made to +1 408 555 1234 or 408 555 1234 is modified and rerouted to 51234 if the call is made while the dual-mode device is registered to Unified CM.

- Ensure that all inbound calls to the enterprise destined for dual-mode devices have the calling number and/or caller ID prefixed with appropriate digits so that missed, placed, and received call history lists are in full +E.164 formats. This will allow dual-mode device users to dial from call history lists without the need for editing the dial string. Instead, users will be able to select a number from the call history list to redial, whether connected to the enterprise or not. For example, if an incoming call from inside the enterprise originates from 51234 to a dual-mode user's enterprise number and the call goes unanswered, Unified CM should be configured to manipulate the calling number so that the resulting entry within the history list of the dual-mode device shows either 408 555 1234 or +1 408 555 1234. This number can be dialed whether the dual-mode device is connected to the enterprise or just to the mobile voice network without the need for further manipulation.

The one exception to normalized dialing for dual-mode devices is for scenarios in which some enterprise extensions or phones are reachable only internally (that is, they have no externally reachable corresponding DID number). In these situations, non-externally reachable numbers can be dialed (manually or from contacts) using abbreviated formats. Because these numbers will never be available externally and can be dialed only from inside the enterprise, some sort of enterprise-only indication should be made when storing these numbers in the contact list. Further, incoming calls from these internal-only numbers should not have the calling number modified for call history lists because these numbers may be called only inside the enterprise. Instead, calls from these extensions should be listed in all call history lists without modification so that the abbreviated dial strings can be successfully dialed only while the device is connected to the enterprise and registered to Unified CM.

Mobile client devices that do not have cellular voice radios, such as tablets, are dependent exclusively on enterprise connectivity and enterprise voice and video telephony or cloud-based collaboration services.

Emergency Services and Dialing Considerations

Mobile client devices do present a slight challenge when it comes to making calls to emergency service numbers such as 911, 999, and 112. Because the mobile client devices may be located inside or outside the enterprise, providing location indication of a device and its user in the event of an emergency must be considered. Dual-mode mobile devices with cellular voice radios receive location services from their provider networks, and these location services are always available when the device is connected and typically able to pinpoint locations far more precisely than enterprise wireless networks; therefore, Cisco recommends that dual-mode device users rely on the mobile voice network for making emergency calls and determining device and user location. To ensure that Cisco dual-mode client devices rely exclusively on the mobile provider voice network for emergency and location services, these clients force all calls made to numbers configured in the Emergency Numbers field on the mobile client device configuration page to route over the mobile voice network. Further, dual-mode phone users should be advised to make all emergency calls over the mobile voice network rather than the enterprise network.

While making emergency calls over WLANs or mobile data networks is not recommended, mobile devices that do not have cellular voice radios are capable of making calls only through these data interfaces. Mobile devices that do not have cellular voice radios should not be relied upon for making emergency calls.

Enterprise Caller ID

When mobile client devices are connected to the enterprise and registered to Unified CM (either through the mobile data network or a WLAN), all calls made with the enterprise line over the WLAN or mobile data network will be routed with the user's enterprise number as caller ID. This ensures that returned calls made from call history lists at the far-end are always routed through the enterprise because the return call is to the user's enterprise number. If a dual-mode mobile device user has been enabled for

Cisco Unified Mobility, and Single Number Reach is turned on for the mobile phone number, return calls to the enterprise number would also be extended to the dual-mode device through the PSTN whenever it is not connected to the enterprise.

Mid-Call Features

When mobile client devices are connected to the enterprise and registered to Unified CM as enterprise endpoints, they are able to invoke call processing supplementary services such as hold, resume, transfer, and conference, using SIP call signaling methods as supported by Unified CM. Just as with any IP phone or client registered to Unified CM, these devices are able to leverage enterprise media resources such as music on hold (MoH), conference bridges, media termination points, and transcoders.

External Call Routing

When dual-mode mobile client devices are not connected to the enterprise and/or not registered to Unified CM, they may make and receive calls only over the mobile voice network. For this reason, Unified CM has no visibility into any calls being made or received at the dual-mode mobile device while it is unregistered. The mobile number is the caller ID being sent to the network when calls are made from dual-mode phones not connected to the enterprise. This will likely result in unanswered calls being made directly back to the dual-mode device's mobile number instead of being routed back through the enterprise.

If the dual-mode mobile client device is integrated with Cisco Unified Mobility, enterprise two-stage dialing services may be leveraged for making calls through the enterprise network even when the dual-mode device is outside the enterprise and not registered to Unified CM. Unified Mobility two-stage dialing is done using either Mobile Voice Access or Enterprise Feature Access and requires the user to dial an enterprise system access DID number and enter credentials prior to dialing the number they are calling. For more information on Unified Mobility two-stage dialing features, see Mobile Voice Access and Enterprise Feature Access, page 21-59.

Likewise, if the dual-mode phone is integrated with Unified Mobility, a user can also receive incoming calls to their enterprise number at the mobile number through Single Number Reach; can invoke mid-call features using DTMF key sequences including hold, resume, transfer, and conference; and can perform desk phone pickup to move an active call from the mobile phone to the enterprise desk phone.

Remote Secure Enterprise Connectivity

Mobile client devices can utilize the IP telephony infrastructure for enterprise voice and video over IP calling and other collaboration services, even when not inside the enterprise, provided they have a secure connection back to the enterprise in order to register the client with Unified CM and to access other collaboration applications and services. Remote secure connectivity for these devices requires the use of the Cisco AnyConnect mobile client VPN solution or the VPN-less Cisco Expressway mobile and remote access feature in order to secure the client connection over the Internet.

Voice and video quality and user experience for remotely attached mobile client devices will vary depending on the nature of the Internet-based network connection. Cisco cannot guarantee acceptable voice and video quality nor successful connectivity for these types of client connections. Care should be taken when relying on these types of connections for business-critical communications. In the case of dual-mode devices with unreliable or low-bandwidth Internet connections, users with dual-mode devices should be advised to make calls over the mobile voice network if connectivity is available rather than relying on the remote enterprise telephony infrastructure.

Additional Services and Features

In addition to call processing or call control services, Cisco mobile clients and devices are capable of providing the additional features and services described in this section.

Dual-Mode Call Handoff

One very important aspect of dual-mode device deployments is call preservation as a user moves in and out of the enterprise or as the device connects to and disconnects from the enterprise network and network connectivity changes from the cellular voice radio to the WLAN radio, and vice versa. Because dual-mode phone users are often mobile, it is important to maintain any active call as a dual-mode user moves in or out of the enterprise. For this reason, dual-mode client devices and the underlying enterprise telephony network should be capable of some form of call handoff.

There are two types of call handoff that should be accommodated by both the dual-mode client and the underlying IP telephony infrastructure:

- Hand-out

  Call hand-out refers to the movement of an active call from the WLAN or mobile data network interface of the dual-mode phone to the cellular voice interface of the dual-mode phone. This requires the call to be handed out from the enterprise IP network to the mobile voice network through the enterprise PSTN gateway.

- Hand-in

  Call hand-in refers to the movement of an active call from the cellular voice interface of the dual-mode phone to the WLAN or mobile data network interface of the dual-mode phone. This requires the call to be handed in from the mobile voice network to the enterprise IP network through the enterprise PSTN gateway.

The handoff behavior of a dual-mode phone depends on the nature of the dual-mode client and its particular capabilities. Dual-mode client handoff may be invoked manually by the user or automatically based on network conditions. In manual handoff scenarios, the dual-mode users are responsible for engaging and completing the handoff operation based on their location and needs. With automatic handoff, the mobile client monitors the WLAN signal and makes handoff decision based on strengthening or weakening of the WLAN signal at the client. Hand-out is engaged in the case of a weakening WLAN signal, while hand-in is engaged in the case of a strengthening WLAN signal. Automatic handoff depends on the mobile device to provide capabilities for monitoring WLAN signal strength.

Handoff operations are critical for maximizing utilization of the enterprise IP telephony infrastructure for phone calls. These operations are also necessary for providing voice continuity and good user experience so that users do not have to hang up the original call and make another call to replace it.

XMPP-Based IM and Presence

Some mobile clients are capable of providing enterprise instant messaging (IM) and presence services based on the Extensible Messaging and Presence Protocol (XMPP), through integration to an on-premises or off-premises application server or service. In either case, the IM and presence capabilities of these mobile clients enable the following:

- Adding users to contact or buddy lists
- Setting and propagating user presence and availability status
- Reception of presence status for a buddy or contact
- Creating and sending of instant messaging (IM) or text messages
- Reception of IM or text messages

While IM and presence are not required functionality for mobile clients, they do enable users to make their availability status visible to contacts and to view the availability status of contacts, thus improving productivity. Further, users can send enterprise-based IM messages rather than incurring costs for mobile Short Message Service (SMS) messages.

Corporate Directory Access

Mobile clients and devices are capable of accessing the enterprise directory for contact lookups. Enterprise directory access is enabled using either:

- Lightweight Directory Access Protocol (LDAP) for communication between the clients and a compatible LDAP directory
- REST-based (HTTPS) communications between the clients and the User Data Services (UDS) API, which provides a set of operations that enable authenticated access to user contact information stored within the end-user database of the Unified CM cluster

  UDS-to-LDAP Proxy can also be used for contact searches. When enabled, contact searches are still handled by UDS but are proxied to the corporate LDAP directory, with UDS relaying results back to the mobile client. This enables mobile clients to search a corporate directory that exceeds the number of users supported within Unified CM.

While corporate directory access is not a required feature for mobile devices and clients, it does provide a superior user experience for mobile users when they are able to access corporate directory information from their mobile device.

Enterprise Voicemail Services

Many mobile clients and devices are also capable of accessing enterprise voicemail services. Cisco mobile clients are capable of receiving enterprise message waiting indication whenever an unread voicemail is in the user's enterprise voicemail box and the mobile device is attached to the enterprise network. Further, mobile clients can be used to retrieve enterprise voicemail messages. Typically enterprise voicemail messages are retrieved when the user dials the voicemail system number and navigates to their voicemail box after providing required credentials. However, Cisco Jabber mobile clients provide the ability to retrieve voicemail messages from the voicemail box by downloading and displaying a list of all messages in the voicemail box and then by selecting individual messages to be downloaded to the mobile device for listening. This is sometimes referred to as visual voicemail. Both the mobile client and the enterprise voicemail system must be capable of providing and receiving message waiting indication (MWI), voicemail message information, and downloads of the messages over the network. Cisco Unity Connection supports visual voicemail through REST (HTTPS) and provides MWI, voicemail lists, and message downloads.

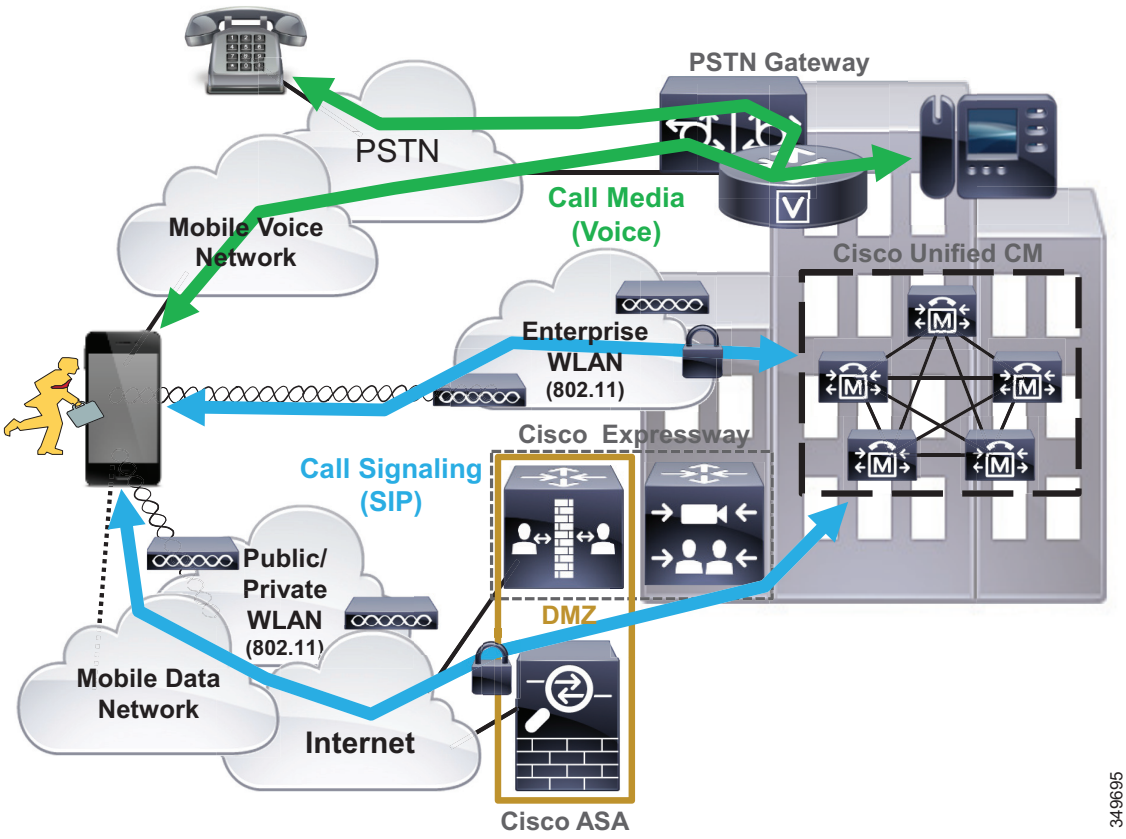
Dial Via Office

Dial via office (DVO) functionality provides automated enterprise dialing capabilities that enable dual-mode mobile devices to initiate calls through the enterprise telephony infrastructure. Deploying DVO calling provides the following benefits to the enterprise:

- Cost savings for calls to international and (possibly) long distance destinations as compared to direct-dialed cellular calls. Note that, in cases of mobile data traversal, mobile data costs must also be considered.
- Ability to dial internal enterprise numbers. Because DVO calls are made using the enterprise line, non-DID or internal-only enterprise extensions are reachable.
- Mobile phone number masking. For DVO calls, the system sends the user's enterprise number as caller ID, and not the mobile phone number.
- Centralized enterprise call detail records (CDRs) and call logs. Because DVO calls are made through the enterprise telephony infrastructure, administrators have complete visibility to these calls even though they traverse the PSTN and mobile voice network.
- Enterprise call anchoring. DVO calls are anchored in the enterprise, thus enabling users to leverage Cisco Unified Mobility DTMF-based mid-call features and desk phone pickup.

Dual-mode mobile devices running the Cisco Jabber client are able to make DVO calls using the Unified CM telephony infrastructure and enterprise PSTN gateway to make calls using the enterprise line. However, unlike voice over IP (VoIP) calling where voice media traverses the IP network, this functionality is facilitated by SIP signaling between the client and Unified CM over an IP connection (WLAN or mobile data) and voice media between the mobile device and the mobile voice network and PSTN, as shown in Figure 21-28.

Figure 21-28 Cisco Dial Via Office Architecture



Note For DVO calls, all voice or media from the user's mobile phone will always travel through the mobile voice network, PSTN, and enterprise PSTN gateway. Media never traverses the IP data connection to the enterprise. The mobile data network connection is used only for call signaling traffic and other application interactions.

For details on dial via office as implemented for Cisco Jabber clients, refer to Cisco Jabber Dial Via Office for Dual-Mode Devices, page 21-96.

Simplified Configuration for Mobile Client Users

Cisco mobile clients provides a streamlined configuration method for simplifying first-time end-user client configuration at the mobile client device. This configuration method relies on RFC 2782 standard Domain Name Service records (DNS SRV) within the corporate DNS server to automatically discover collaboration services on the network. DNS SRV records direct the mobile client to appropriate

application servers for call control and IM and presence services. This configuration and provisioning method alleviates the need for the user to manually configure the XMPP IM and presence server and voice and video call control server or TFTP server host name or IP address. Instead the user simply enters their user ID and domain name, and the client application automatically discovers the available collaboration services and connects to these back-end servers, with the application prompting the user for credentials as appropriate. If no services are discovered or if service discovery operation fails, then the mobile client application reverts to manual configuration mode, requiring users to enter collaboration application server host names or IP addresses and credentials. Multiple DNS SRV records with priority and weighting indication ensure high availability of back-end collaboration application services as well as mobile client distribution across multiple servers providing these services.

Note Mobile client user simplified configuration does not simplify administrative tasks related to client and service configuration and provisioning on the back-end application servers. All administrative tasks to add user accounts, mobile client devices, and services configuration are still required in addition to creating the DNS SRV record or records in the corporate DNS server.

Cisco Bring Your Own Device (BYOD) Infrastructure

Cisco mobile client applications such as Cisco Jabber provide core Unified Communications and collaboration capabilities, including voice, video, and instant messaging to users of mobile devices such as Android and Apple iOS smartphones and tablets. When a Cisco mobile client device is attached to the corporate wireless LAN, the client can be deployed within the Cisco Bring Your Own Device (BYOD) infrastructure.

Because Cisco mobile clients and devices rely on enterprise wireless LAN connectivity or remote secure attachment through VPN or VPN-less connections, they can be deployed within the Cisco Unified Access network and can utilize the identification, security, and policy features and functions delivered by the BYOD infrastructure.

The Cisco BYOD infrastructure provides a range of access use cases or scenarios to address various device ownership and access requirements. The following high-level access use case models should be considered:

- Basic Access — This use case enables basic Internet-only access for guest devices. This use case provides the ability to enable employee-owned personal device network connectivity without providing access to corporate resources.
- Limited Access — This use case enables full access to corporate network resources, but it applies exclusively to corporate-owned devices.
- Enhanced Access — This use case enables granular access to corporate network resources for both corporate-owned devices and employee-owned personal devices based on corporate policies.

Cisco collaboration mobile clients, whether running on corporate or personal devices, usually require access to numerous back-end on-premises enterprise application components for full functionality. For this reason the Limited or Enhanced Access use case scenarios generally apply to applications such as Cisco Jabber for Android or iPhone. The chief difference between these two access models is that with Limited Access, the corporate-owned devices are given full access to corporate network resources. In the case of Enhance Access, not only is the scope expanded to include employee-owned devices, but access to corporate network resources can also be provided in a granular way so that devices and the applications that run on them are able to access only specific resources based on corporate security policies.

In the case of cloud-based collaboration services, Cisco mobile clients and devices connect directly to the cloud through the Internet without the need for enterprise network attachment. In these scenarios, user and mobile devices can be deployed using the Basic Access model because these use cases require only Internet access.

For more information about the Cisco BYOD infrastructure and BYOD access use cases, refer to the BYOD information available at

https://www.cisco.com/c/en/us/solutions/byod-smart-solution/overview.html

When deploying Cisco mobile clients and devices within the Cisco BYOD infrastructure, consider the following high-level design and deployment guidelines:

- The network administrator should strongly consider allowing voice and video-capable clients to attach to the enterprise network in the background (after initial provisioning), without user intervention, to ensure maximum use of the enterprises telephony infrastructure. Specifically, use of certificate-based identity and authentication helps facilitate an excellent user experience by minimizing network connection and authentication delay.
- In scenarios where Cisco mobile clients and devices are able to connect remotely to the enterprise network through a secure VPN or VPN-less connection:
  - The network administrator should weigh the corporate security policy against the need for seamless secure connectivity without user intervention in order to maximize utilization of the enterprise telephony infrastructure. The use of certificate-based authentication and enforcement of a device pin-lock policy provides seamless attachment without user intervention and functionality similar to two-factor authentication because the end user must possess the device and know the pin- lock to access the network. If two-factor authentication is mandated, then user intervention will be required in order for the device to attach remotely to the enterprise.
  - It is important for the infrastructure firewall configuration to allow all required client application network traffic to access the enterprise network. Failure to provide an appropriate access solution or to open access to appropriate ports and protocols at the corporate firewall could result in an inability of the Cisco mobile clients or devices to register to on-premises Cisco call control for voice and video telephony services and/or the loss of other client features such as enterprise directory access or enterprise visual voicemail.
- When enterprise collaboration applications such as Cisco Jabber are installed on employee-owned mobile devices, if the enterprise security policy requires the device to be wiped or reset to factory default settings under certain conditions, device owners should be made aware of the policy and encouraged to backup personal data from their device regularly.
- When deploying Cisco collaboration mobile clients and devices, it is important for the underlying network infrastructure from end-to-end to support the necessary QoS classes of service, including priority queuing for voice media and dedicated video and signaling bandwidth, to ensure the quality of client application voice and video calls and appropriate behavior of all features.

Deployment Considerations for Cisco Mobile Clients and Devices

This section discusses deployment considerations the following Cisco mobile clients and devices:

- Cisco Jabber for Android and Apple iOS, page 21-90
- Cisco Spark, page 21-108
- Cisco WebEx Meetings, page 21-108
- Cisco AnyConnect Mobile Client, page 21-109

Cisco Jabber for Android and Apple iOS

This section describes characteristics and deployment considerations for Cisco Jabber.

Cisco Jabber mobile clients are available for Android and Apple iOS mobile devices, including iPad and iPhone. Once the client application is downloaded from the appropriate store or market (Apple Application Store or Google Play) and installed on the Apple iOS or Android device, it can connect to the enterprise network and register to Unified CM as a SIP enterprise phone.

To provide registration and call control services to the Cisco Jabber mobile client, the device must be configured within Unified CM as a **Cisco Dual Mode for Android** or **iPhone**, or **Cisco Jabber for Tablet** device type. Next, the mobile device must be configured to access the enterprise WLAN for connectivity based on the enterprise WLAN infrastructure and security policies. Alternatively the mobile device can be connected to the enterprise network via the mobile data network or over non-enterprise WLANs. Once the mobile device has been configured to access the enterprise network, when the Cisco Jabber client is launched, it will register the device to Unified CM. To integrate to Unified Mobility and to leverage handoff functionality, the mobile number of the Android or iPhone smartphone must be configured as a mobility identity associated to the Cisco Dual Mode for Android or iPhone device within Unified CM.

The Cisco Jabber client is supported on the following devices:

- Android

  Various models of Android phones and tablets. (Consult the release notes referenced below for specific device and firmware support information.) These devices must be running a minimum firmware version of 4.1(2), although later versions of Android firmware may be required. The WLAN interfaces of most Android devices support 802.11a, 802.11b, 802.11g, 802.11n, and 802.11ac network connectivity.

- Apple iOS

  Various Apple iOS devices including iPhone and iPad. (Consult the release notes referenced below for specific device and firmware support information.) These devices must be running a minimum iOS version of 10.3. The WLAN interfaces of most Apple iOS devices support 802.11a, 802.11b, 802.11g, and 802.11n network connectivity. Some newer Apple devices support 802.11ac.

For details on the latest specific device and firmware version support, refer to the product release notes for:

- Android

  https://www.cisco.com/c/en/us/support/unified-communications/jabber-android/products-release-notes-list.html

- iPhone and iPad

  https://www.cisco.com/c/en/us/support/customer-collaboration/jabber-iphone-ipad/products-release-notes-list.html

The Cisco Jabber for Android, iPad, and iPhone clients not only provide voice and video over IP phone services but also provide XMPP-based enterprise instant messaging (IM) and presence, corporate contact and directory services when configured to access the enterprise contact source, and enterprise voicemail message waiting indication (MWI) and visual voicemail when integrated to Cisco Unity Connection.

The Cisco Jabber clients running on smartphones (Android and iPhone) are capable of performing only manual hand-out as described in the section on Cisco Jabber Dual-Mode Handoff, page 21-93.

For more information about the Cisco Jabber Android and Apple iOS clients, additional feature details, and supported hardware and software versions, refer to the Cisco Jabber documentation for:

- Android

  https://www.cisco.com/c/en/us/support/unified-communications/jabber-android/tsd-products-support-series-home.html

- iPhone and iPad

  https://www.cisco.com/c/en/us/support/customer-collaboration/jabber-iphone-ipad/tsd-products-support-series-home.html

Cisco Jabber Service Discovery

As indicated previously, Cisco mobile clients such as Jabber are able to discover available collaboration services by relying on DNS lookups and DNS SRV service record resolution. When service discovery is properly configured, the user needs to enter only their user name and domain, and the client will automatically discover and connect to available collaboration services.

As shown in Figure 21-29, during initial client configuration or in the case of network connection changes, Jabber discovers collaboration services by querying DNS for the following SRV records:

- _cisco_uds._tcp.<*domain*>

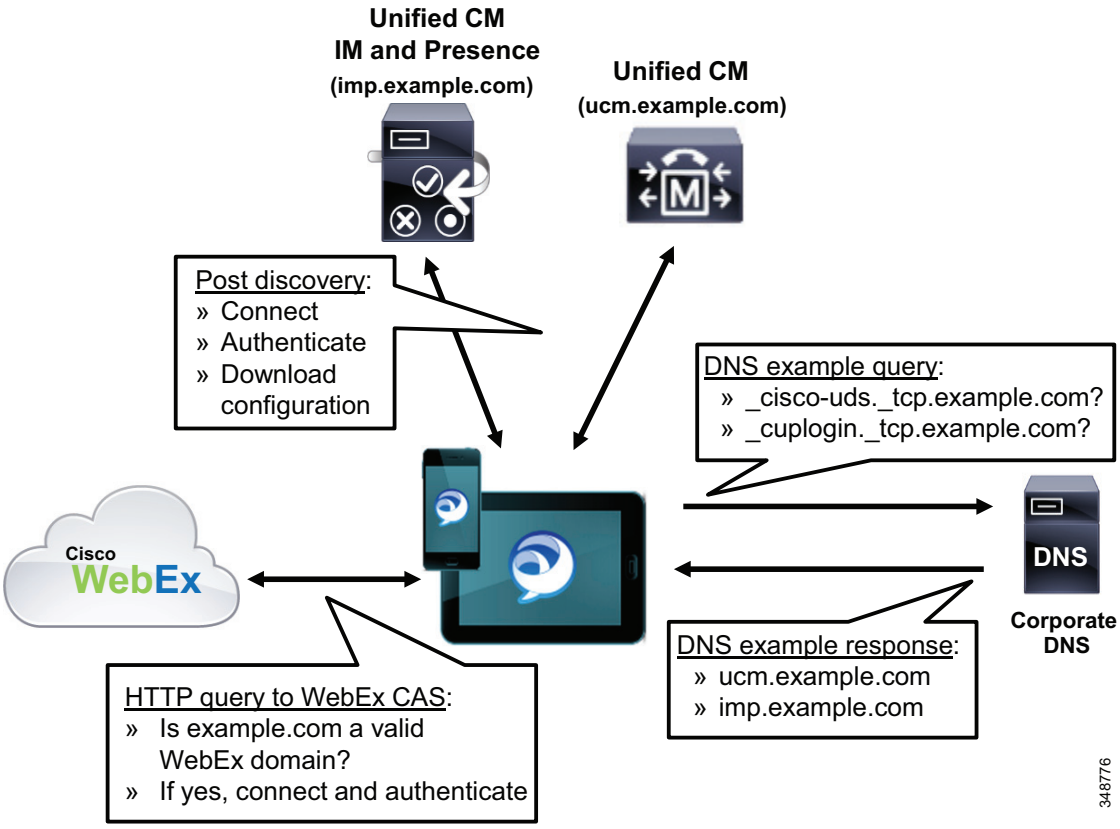
  SRV record or records of this type are added to the enterprise DNS server when Jabber is deployed in phone-only mode enabling voice and video over IP calling or in full UC mode enabling both voice and video calling well as IM and presence. If the query for this record is resolved by DNS, Cisco Jabber connects to Unified CM, determines the authenticator, and locates available services.

- _cuplogin._tcp.<*domain*>

  SRV record or records of this type are added to the enterprise DNS server when Jabber is deployed in IM-only mode enabling XMPP-based IM and presence. If the query for this record is resolved by DNS, Cisco Jabber connects to Unified CM IM and Presence and authenticates.

In the case of hybrid deployments with Cisco WebEx Messenger, during initial configuration and on network connection changes, the client also issues an HTTP query to a central authentication service (CAS) URL for Cisco WebEx Messenger service to determine if the domain is a valid WebEx domain. When the client receives positive confirmation to the HTTP query that a valid WebEx domain has been entered, the client then connects to and authenticates with the WebEx Messenger service and retrieves client configuration and information on available UC services as configured in the Cisco WebEx Org Admin.

Figure 21-29 Cisco Jabber Service Discovery



While the UDS service runs on all nodes in the Unified CM cluster, when configuring DNS SRV records for Unified CM UDS service, administrators should configure records for resolution to Unified CM subscriber nodes only. This ensures that client interaction with the UDS service avoids the publisher node and instead distributes the load across call processing nodes within the cluster.

In deployments where service discovery is not configured or reliance on DNS is not possible, the Jabber client will revert to manual configuration, requiring the user to enter authenticator and service node IP addresses. Manually configured IP addresses are cached by the Jabber client for use on subsequent connections.

Once service discovery or manual configuration is complete, Jabber must authenticate and download a service profile and/or the jabber-config.xml file (if available), which directs the client to additional back-end application services such as voicemail and directory and enables appropriate configuration.

Cisco Jabber Corporate Directory Access

Cisco Jabber mobile clients rely on various methods for accessing enterprise contact information. In addition to local device contacts and contacts previously added to the Jabber buddy list, Jabber mobile clients are also able to access corporate directory services using the following methods:

- Cisco Directory Integration (CDI)

  The CDI method of corporate directory access relies on LDAP communication between the Jabber client and supported LDAP compliant directories such as Microsoft Active Directory and OpenLDAP. CDI is the default method of directory integration for on-premises Jabber clients.

- Unified CM User Data Services (UDS)

  The UDS method of corporate directory access relies on HTTP communication between the Jabber client and Unified CM UDS services running on each Unified CM node.

- Unified CM UDS-to-LDAP Proxy

  This method of corporate directory access relies on the Unified CM UDS service resolving or proxying directory searches against the corporate LDAP directory rather than using the local user directory. UDS-to-LDAP proxy allows Jabber users to search against the entire corporate directory rather than being limited by the local Unified CM cluster end-user database.

The jabber-config.xml file is used to configure the directory integration method for Jabber clients as well as to configure certain directory related settings for Jabber clients.

We recommend using the CDI method of directory access for on-premises clients.

When Jabber clients connect remotely using Expressway mobile and remote access, only UDS methods of directory access (local Unified CM database or UDS-to-LDAP proxy) are supported. Consider enabling UDS-to-LDAP proxy when the corporate directory size exceeds the local Unified CM directory size (greater than 160,000 users), to enable mobile client users to search the entire directory.

Cisco Jabber Dual-Mode Handoff

To properly deploy Cisco dual-mode clients such as Cisco Jabber, it is important to understand the nature of handoff operations within the client. The handoff method used by the Cisco Jabber dual-mode client depends on the **Transfer to Mobile Network** setting on the Cisco Dual Mode for iPhone or Cisco Dual Mode for Android device configuration page.

There are two methods of handoff, depending on the Transfer to Mobile Network setting:

- Mobility Softkey Method of Hand-Out, page 21-93

  With this method the Transfer to Mobile Network setting should be set to **Use Mobility Softkey (user receives call)**. In this type of handoff, the Unified CM system generates a call over the PSTN to the user's mobile number.

- Handoff Number Method of Hand-Out, page 21-94

  With this method the Transfer to Mobile Network setting should be set to **Use HandoffDN Feature (user places call)**. In this type of handoff, the mobile client generates a call over the mobile voice network to the handoff number configured within the Unified CM system.

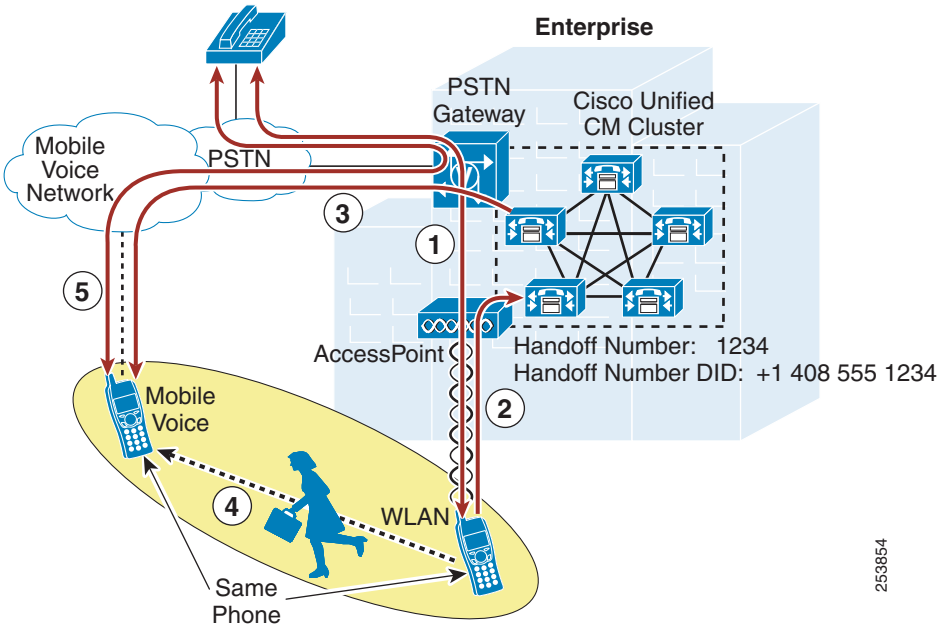
Note Handoff capabilities apply only to dual-mode smartphones. This functionality is not supported on devices without cellular voice radios, such as the Samsung Galaxy Note Pro.

Mobility Softkey Method of Hand-Out

The operation depicted in Figure 21-30 is of an active call on an iPhone or Android dual-mode device within the enterprise being moved manually from the WLAN interface to the mobile voice network or cellular interface of the device through the enterprise PSTN gateway. As shown, there is an existing call between the mobile client device associated to the enterprise WLAN and registered to Unified CM, and a phone on the PSTN network (step 1). Because this is a manual process, the user must select the Use Mobile Network button from the in-call menu within the Cisco Jabber client, which signals to Unified CM the intention to hand-out the call (step 2). Next Unified CM generates a call to the configured mobility identity number corresponding to this mobile device through the enterprise PSTN gateway (step 3). This call to the mobility identity is made to the mobile voice network or cellular interface of the iPhone or Android device. The user can now move out of the enterprise and away from WLAN network coverage (step 4). In the meantime, the inbound call from Unified CM is received at the mobile voice network interface, and the user must answer the call manually to complete the hand-out.

Once the inbound call on the cellular interface is answered, the RTP stream that was traversing the WLAN is redirected to the PSTN gateway, and the call continues uninterrupted between the mobile client device and the original PSTN phone, with the call anchored in the enterprise gateway (step 5).

Figure 21-30 Cisco Jabber Dual-Mode Hand-Out (WLAN-to-Mobile Voice Network): Mobility Softkey Method

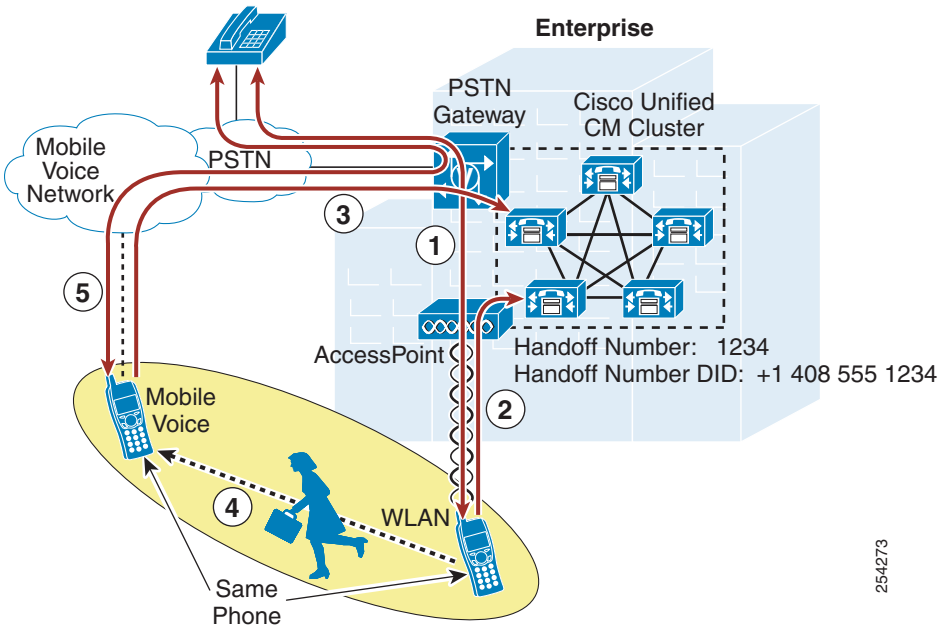


Handoff Number Method of Hand-Out

Figure 21-31 illustrates the same hand-out operation as in Figure 21-30, where an active call on an iPhone dual-mode phone within the enterprise is moved manually from the WLAN interface to the mobile voice network or cellular interface of the device through the enterprise PSTN gateway. However, in this case the Handoff Number method of hand-out is used.

As shown in Figure 21-31, there is an existing call between the dual-mode device associated to the enterprise WLAN and registered to Unified CM, and a phone on the PSTN network (step 1). Because this is a manual process, the user must select the Use Mobile Network button from the in-call menu within the Cisco Jabber dual-mode client, which signals to Unified CM the intention to hand-out the call (step 2). Next the Cisco Jabber client automatically generates a call through the cellular interface over the mobile voice network to the configured Handoff Number within the Unified CM system (step 3). The user can now move out of the enterprise and away from WLAN network coverage (step 4). In the meantime, the inbound call from the Cisco Jabber client is received by Unified CM. Assuming the inbound calling number matches the user's configured mobility identity, the RTP stream that was traversing the WLAN is redirected to the PSTN gateway, and the call continues uninterrupted between the Cisco Jabber mobile client and the original PSTN phone, with the call anchored in the enterprise gateway (step 5).

Figure 21-31 Cisco Jabber Dual-Mode Hand-Out: Handoff Number Method



Note The Handoff Number method of hand-out requires Unified CM to receive an inbound calling number from the PSTN network that matches the mobility identity number configured under the Cisco Dual Mode device attempting the hand-out. If the caller ID is not sent by the dual-mode device, if the PSTN provider does not send the inbound caller ID to the enterprise, or if the inbound caller ID does not match the user's configured mobility identity, the hand-out operation will fail.

Note Cisco Jabber dual-mode clients do not support hand-in. In scenarios where an in-progress call is active between the dual-mode mobile voice network or cellular interface and an enterprise phone (or a PSTN phone with the call anchored in the enterprise gateway), the only way to move the call to the WLAN interface of the dual-mode device is to hang up the call and redial once the dual-mode client has connected to the enterprise network and registered to Unified CM.

WLAN Design Considerations for Cisco Jabber Mobile Clients

Consider the following WLAN guidelines when deploying Cisco Jabber mobile clients:

- Whenever possible, ensure that Cisco Jabber mobile clients roam on the WLAN only at Layer 2 so that the same IP address can be used on the WLAN interface of the device. In Layer 3 roaming scenarios where subnet boundaries are crossed due to device IP address changes, calls will be dropped.
- Deploy Cisco Jabber mobile clients on WLAN networks where the same SSID is used across all APs. Roaming between APs is much slower if SSIDs are different.
- Ensure all APs in the WLAN broadcast their SSID(s). If the SSID is not broadcast by the AP, the user may be prompted by the device to join other Wi-Fi networks or the device may automatically join other Wi-Fi networks. When this occurs the call is interrupted.

- Whenever possible, deploy Cisco Jabber mobile clients on the 5 GHz WLAN band (802.11a/n/ac). 5 GHz WLANs provide better throughput and less interference for voice and video calls.

Cisco Jabber Dial Via Office for Dual-Mode Devices

The Unified CM administrator can enable or disable dial via office (DVO) calling for each dual-mode device by using the Product Specific Configuration Layout section of the Cisco Dual Mode for iPhone or Android device configuration page. Once DVO is enabled, the user can turn on DVO using the Calling Options setting within the Cisco Jabber application. It is important to note that the DVO calling options dictate not only the outbound calling method used by the Jabber client but also the inbound calling method. Table 21-4 shows the various calling options and the corresponding outbound and inbound calling method based on the type of network connectivity.

Table 21-4 Inbound and Outbound Calling Method with Cisco Jabber Dial Via Office Calling Options

Device IP Connection	Cisco Jabber DVO Calling Options					
	Autoselect		Mobile Voice Network		Voice over IP	
	Outbound Call	Inbound Call	Outbound Call	Inbound Call	Outbound Call	Inbound Call
802.11 WLAN (Corporate/enterprise)	Voice over IP	Voice over IP	Dial via office	Single Number Reach	Voice over IP	Voice over IP
802.11 WLAN (Non-corporate/enterprise)						
Mobile Data	Dial via office	Single Number Reach				
No IP	Outbound call: Native cellular Inbound call: Single Number Reach					

The default calling option when DVO is first enabled is Autoselect, which results in voice over IP (VoIP) for both inbound and outbound Cisco Jabber calling when the device is connected over an 802.11 WLAN, while DVO will be used for outbound calling and Single Number Reach will be used for inbound calling when the device is connected over the mobile data network.

In all cases, calls made to emergency numbers configured in the Emergency Numbers field on the mobile client device configuration within Unified CM will be dialed directly over the cellular network regardless of the calling option selected.

Note The Dial via Office calling feature applies only to dual-mode smartphones. This functionality is not supported on tablets such as the Apple iPad because there is no cellular voice radio on those devices.

When dial via office is enabled for Cisco Jabber clients, as with Single Number Reach, the mobile voicemail avoidance or single enterprise voicemail box feature of Cisco Unified Mobility is engaged. In the case of dial via office, this voicemail avoidance feature ensures that, given a failure in the network path or some other communication error during a DVO call setup, the called user does not end up in the calling user's voicemail box. Typically the User Control method of voicemail avoidance provides the best overall user experience because, if a DVO call leg inadvertently ends up being answered by a voicemail system, the call leg will be disconnected when a DTMF tone is not received by Unified CM, and the DVO call will be cleared. When Cisco Jabber users are enabled for the User Control method of mobile voicemail avoidance, they should be reminded that they must press a button on the mobile device key pad when receiving a mobility call at the client device. Failure to do so will result in call setup failure.

Note Because the User Control method of mobile voicemail avoidance is completely dependent on successful relay of the DTMF tone from the mobile device over the PSTN connection and PSTN gateway and out-of-band to Unified CM, failure to propagate inbound DTMF from the PSTN to Unified CM results in a disconnect of all enterprise calls made (dial via office reverse) or received (single number reach) by the mobile device. If DTMF cannot be effectively relayed from the PSTN to Unified CM, then the Timer Control mobile voicemail avoidance method should be used instead.

For more information about the single enterprise voicemail box voicemail avoidance feature, see Mobile Voicemail Avoidance with Single Enterprise Voicemail Box, page 21-55.

Dial Via Office Calling Option Use Cases

When deploying dial via office, consider the following Cisco Jabber client calling option user profiles:

- Autoselect

  The typical user profile for Autoselect is a user that is mobile both within and outside the office. For this user profile, Autoselect provides potential least cost routing by taking advantage of VoIP when 802.11 WLAN connectivity is available and falls back to mobile voice and data network (DVO and Single Number Reach) when WLAN connectivity is not available.

- Mobile Voice Network

  The typical user profile for the Mobile Voice Network calling option is a highly mobile user that almost never has WLAN coverage and whose mobile data connectivity does not provide acceptable throughput and reliability to ensure good voice quality and reliable calling over IP connections.

- Voice over IP

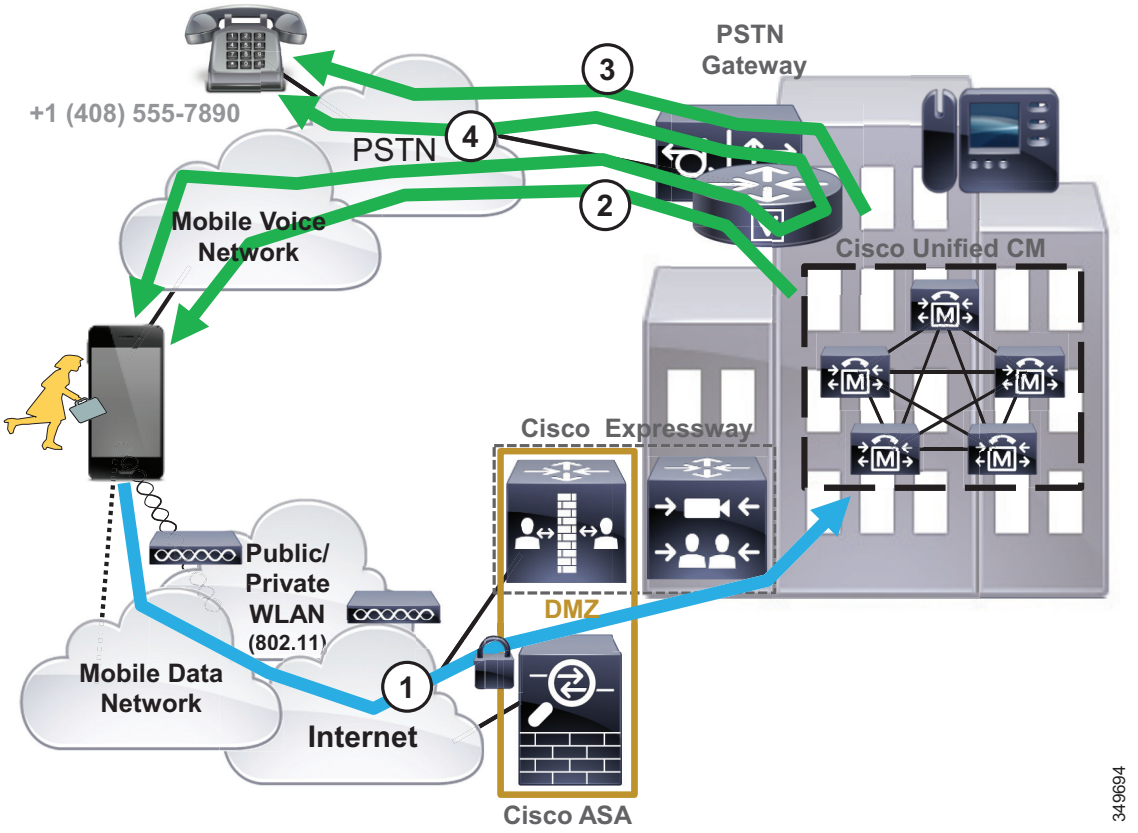
  The typical user profile for the Voice over IP calling option is a user that is mobile within the office (home or enterprise) but for whom enterprise calling is not typically required outside the enterprise. Additionally, with this user profile, mobile voice and data costs are usually an important consideration for both corporate-paid and employee-paid mobile voice and data service.

Dial Via Office Reverse

Cisco Jabber clients support dial via office reverse (DVO-R). With this method of DVO, the call setup is facilitated by an inbound call from the Unified CM system to the user's configured mobility identity or mobile phone number.

Figure 21-32 illustrates a DVO-R call flow. In this example, a Cisco Jabber user wishes to dial a PSTN phone at +1 408 555-7890. The user dials the number or selects the number from the contact list from within the Cisco Jabber client, which generates a SIP call setup request over the IP connection to the enterprise and Unified CM (step 1). Based on the call setup request, Unified CM generates a reverse call back to the user's configured mobility identity (mobile phone number) using the enterprise PSTN gateway (step 2). Once the incoming call from Unified CM is answered at the mobile device, a call is extended to the number the user called or selected (step 3; in this case, +1 408-555-7890). Once the call is answered at the far end, the media path is connected and the call is anchored through the enterprise PSTN gateway (step 4). Because the call is now anchored in the enterprise gateway, the user has the ability at any point during this call to use the Unified Mobility desk phone pickup feature as well as to invoke Unified Mobility DTMF-based mid-call features.

Figure 21-32 Cisco Jabber Dial Via Office Reverse



Note The call flow shown in Figure 21-32 assumes that Cisco Jabber is registered to Unified CM, that DVO is enabled for the user, and that the client calling option setting is either Mobile Voice Network or Autoselect. If the client setting is Autoselect, the dual-mode device running Cisco Jabber must be IP-connected via the mobile data network. If connected over 802.11 WLAN, then the client would use voice over IP rather than DVO.

By default the DVO-R callback call leg will be extended to the user's mobile device, as shown in Figure 21-32; however; a user may specify an alternate callback number in the DVO Callback Number field within the Cisco Jabber client. By default the DVO Callback Number field is populated with the user's configured mobility identity. If the user configures a different number in this field, the DVO-R callback call leg will be extended to that number. For example, rather than receiving the callback on the mobile phone, the user may wish to direct the callback to their home phone.

Note When invoking DVO-R with an alternate callback number, if the callback call leg from Unified CM is directed to a user-specified alternate number, the call is not anchored in the enterprise. In such cases, users cannot perform desk phone pickup or invoke DTMF-based mid-call features on DVO-R calls using an alternate callback number. In addition, voicemail avoidance does not engage for DVO-R alternate number calls.

Note DVO-R calls are facilitated with en-bloc dialing and therefore overlap sending will not be engaged even for patterns with Allow Overlap Sending enabled.

Mobile Profiles and Dial Via Office Reverse

Cisco Unified CM mobility profiles may be assigned to the mobility identity for mobile client devices. While not required, the mobility profile specifies the caller ID sent by the system during setup of the DVO-R callback call leg to the mobility identity or alternate callback number. The number configured in the Callback Caller ID field of the Dial-via-Office Reverse Callback Configuration section of the mobility profile configuration page is the number sent as caller ID. If no mobility profile is assigned to the mobility identity or if the Callback Caller ID field is left blank, the system will send the configured default Enterprise Feature Access Number.

Note The Mobile Client Calling Option field of the mobility profile has no impact on DVO operation; regardless of the setting, the Cisco Jabber client makes DVO-R calls when enabled for DVO calling. Dial via Office Forward (DVO-F) is not a currently available calling option.

Cisco Jabber Point-to-Point Calling

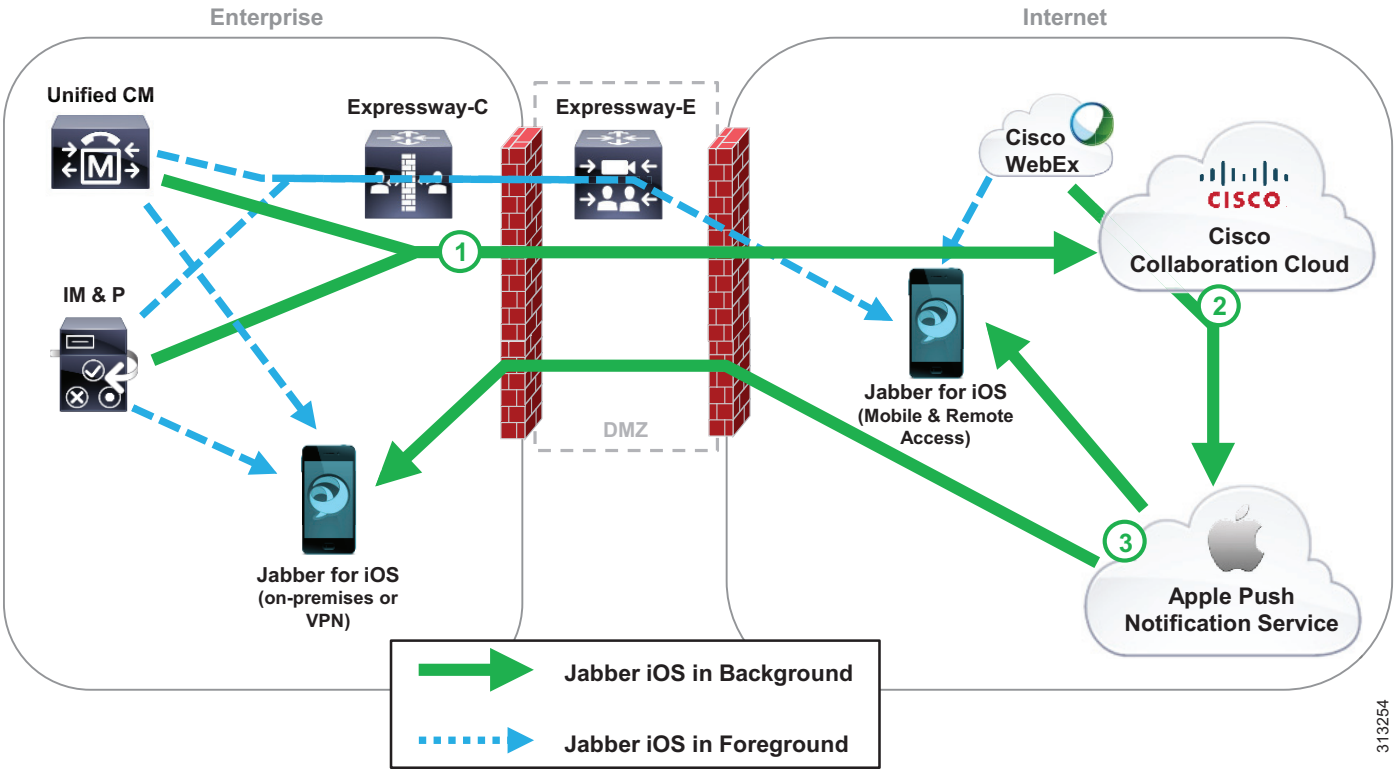
Cisco Jabber mobile clients are capable of providing point-to-point voice and video calling over IP without the need for Unified CM registration. Instead, the Jabber client leverages the Cisco WebEx Messenger cloud service for REST/HTTPS call signaling. Point-to-point call media leverages the RTP protocol with the G.722 codec for call audio and H.264 for call video. With REST point-to-point calling, only a single call per Jabber mobile client is supported, and mid-call supplementary features such as hold, resume, transfer, and conference are not supported.

Apple Push Notification Service (APNs) for Cisco Jabber for iPhone and iPad

When running in the background on mobile devices, previously the Cisco Jabber for iPhone and iPad client relied on periodic direct IP socket keep-alive messages to maintain connectivity for voice and video over IP (VVoIP) and IM and presence services when the client moved to the background. This ensures that the user is notified and the client receives incoming calls and messages. Beginning with Cisco Jabber for iPhone and iPad 11.9 and Cisco Unified CM and IM and Presence Service release 11.5 SU3 and later (as well as with current versions of WebEx Messenger), when the client is running in the background on an Apple iOS device, the client can receive incoming call and message notifications through the Apple Push Notification service (APNs).

Figure 21-33 illustrates the APNs architecture. As indicated by the green arrows, when client notification from Unified CM or the Unified CM IM and Presence Service (or WebEx Messenger) is required, Unified CM and the Unified CM IM and Presence Service (and/or WebEx Messenger service) send outbound HTTPS notification from the enterprise network to the Cisco Collaboration Cloud on the Internet (step 1). The Cisco Collaboration Cloud creates a secure connection to the Apple Push Notification service (APNs) on the Internet and forwards Jabber client notifications to APNs (step 2). In turn, APNs forwards the notification to the Jabber iOS client device (step 3), which previously registered to APNs during the initial provisioning of the Apple device on the carrier network. This notification through APNs triggers an alert to the user. The notification architecture applies whether the Jabber for Apple iOS client is connected on-premises, over VPN, or over Expressway mobile and remote access.

Figure 21-33 Cisco Jabber for Apple iOS and APNs Architecture Overview



As indicated by the blue arrows in Figure 21-33, when the Jabber for Apple iOS client is running in the foreground, notifications are sent directly to the client from Unified CM and the Unified CM IM and Presence Service via SIP and XMPP.

For on-premises Unified CM and Unified CM IM and Presence deployments, APNs for Cisco Jabber for iPhone and iPad clients is enabled on Unified CM by the administrator through the cloud onboarding process. Once enabled, Cisco Jabber for iPhone and iPad clients running in the background will receive call and message notifications through APNs.

Note Because current versions of Apple iOS (including iOS 10 and iOS 11) still support keep-alive messages to maintain connectivity when the Cisco Jabber for iPhone and iPad client is running in the background, enabling APNs on Unified CM is not yet a requirement. However, because Apple is deprecating the direct IP socket method for notifications, APNs will soon be required for sending notifications to Jabber for Apple iOS clients running in the background on the Apple iOS device. Once the current direct IP socket method is removed in a future Apple iOS release, APNs will be the only method for notifying users about an incoming call or message when the Cisco Jabber for iPhone and iPad client is running in the background.

In the case of cloud or hybrid deployments using WebEx Messenger, APNs is enabled within the WebEx cloud by default, and Cisco Jabber for iPhone and iPad 11.9 and later clients will receive IM notifications through APNs while running in the background.

Jabber-to-Jabber calling with WebEx Messenger is not supported with APNs. If you plan to use the Jabber-to-Jabber calling feature with WebEx Messenger, you will need to disable APNs manually with the **< Policies > < Push_Notification_Enabled >** parameter in the jabber-config.xml file. For more information on jabber-config.xml parameters, refer to the latest version of the *Parameters Reference Guide for Cisco Jabber*, available at

https://www.cisco.com/c/en/us/support/customer-collaboration/jabber-iphone-ipad/products-installation-guides-list.html

When end-to-end encryption (AES) policy is set to **enforced** or **optional** with WebEx Messenger, APNs is automatically disabled and the client will receive IM notifications in the usual way when the client is running in the background.

Note The use of APNs for Jabber running in the background applies only to Cisco Jabber for iPhone and iPad clients. Windows, Mac, and Android Jabber clients are not impacted and will continue to receive notifications in the usual way when running in the background.

Cisco Jabber and OAuth with Refresh Login Flow

Beginning with Cisco Jabber 11.9, client authorization and authentication is facilitated using the OAuth 2.0 authorization framework. This provides for faster login and faster re-authentication during launch and network transitions. Prior to Cisco Unified CM 12.0 and Unified CM 11.5(1) SU3, Cisco Jabber used OAuth only when Single-Sign On (SSO) was enabled within the deployment. The OAuth implementation relies on the Unified CM publisher acting as an authorization server responsible for authenticating and then issuing authorization tokens to clients. This token, along with a refresh token, enables the client to request and gain authorization to collaboration services and to quickly renew an expired authorization token using the refresh token. For additional details on the OAuth 2.0 framework, refer to the section on Authorization Framework, page 16-45.

To leverage OAuth for Jabber client authorization and authentication, the **OAuth with Refresh Login Flow** service parameter must be enabled on Cisco Unified CM, Unified CM IM and Presence, and Unity Connection. Likewise, the **Authorize by OAuth token with refresh** setting must be enabled on Expressway-C for Jabber clients to use OAuth over Expressway Mobile and Remote Access.

For more information on deploying OAuth with Cisco Jabber, refer to the latest version of the white paper on *Deploying OAuth with Cisco Collaboration Solution Release 12.0*, available at

https://www.cisco.com/c/en/us/support/unified-communications/jabber-windows/products-installation-guides-list.html

Cisco Jabber and Expressway Mobile and Remote Access

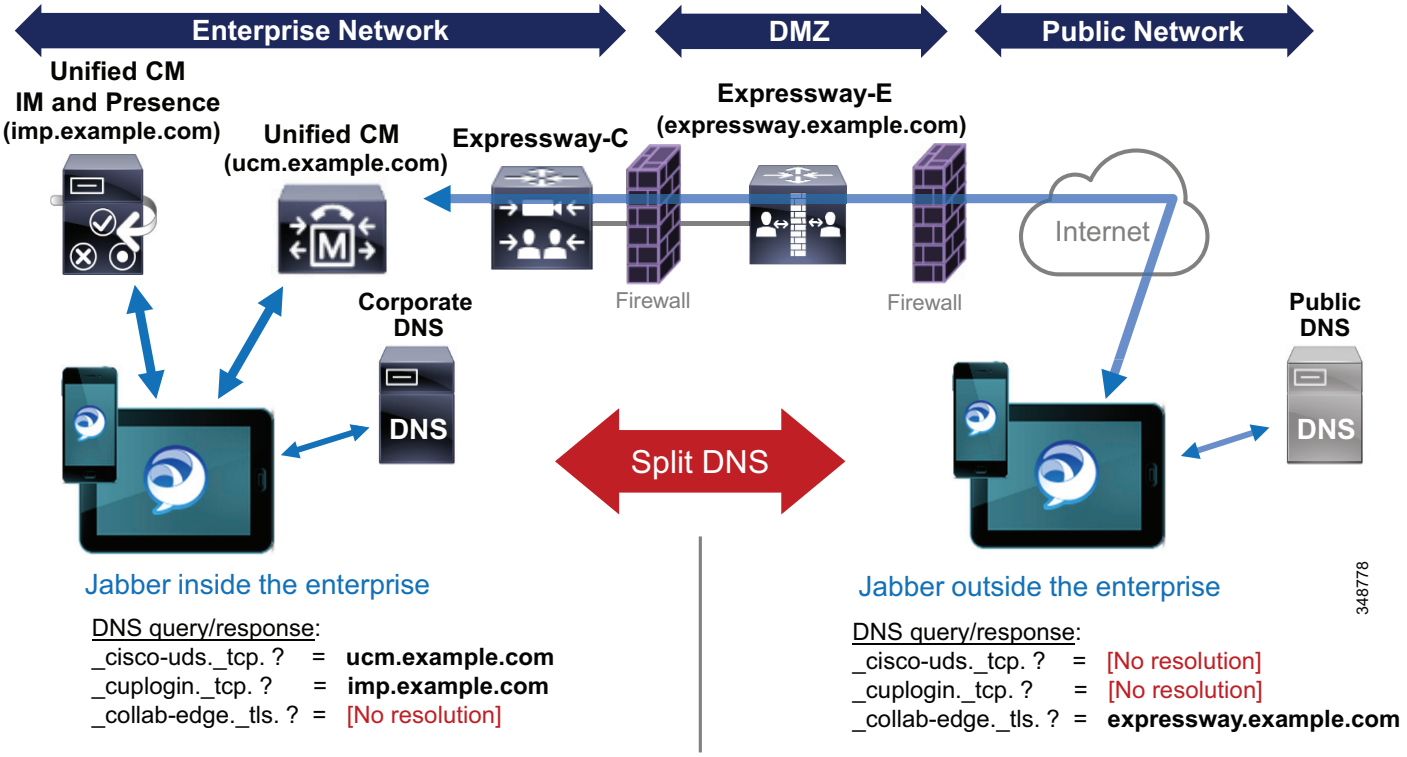
The mobile and remote access feature of the Cisco Expressway solution provides secure firewall traversal for Cisco Jabber, enabling remote Jabber users to access enterprise collaboration applications and services from their mobile devices when outside the enterprise.

All collaboration traffic traversing the Expressway mobile and remote access connection is encrypted, including call media and signaling. The encrypted connection is between the Jabber endpoint and the Expressway-C node inside the enterprise. Traffic between Expressway-C and endpoints or applications inside the enterprise is unencrypted by default. Media and signaling traffic inside the enterprise is encrypted only when the Unified CM cluster is configured as mixed mode with device authentication, SRTP media, and TLS SIP signaling encryption facilitated by security configuration relying on the Unified CM Cisco Certificate Trust List (CTL) Provider and Certificate Authority Proxy Function (CAPF) services.

Jabber determines its location relevant to the enterprise (inside or outside) based on DNS query resolution and a split DNS resolution design whereby the service records for Unified CM (_cisco-uds._tcp) and Unified CM IM and Presence (_cuplogin._tcp) are configured only in the corporate DNS and the service record for Expressway (_collab-edge._tls) is configured only on the public DNS. This split design ensures that corporate DNS resolution points Jabber directly to collaboration services when inside the enterprise and public DNS resolution points Jabber to connect through Expressway. DNS queries are sent by Jabber whenever the network connection of the mobile device changes.

As shown in Figure 21-34, Jabber queries DNS for three SRV service records: _cisco-uds._tcp, _cuplog._tcp, and _collab-edge._tls. When inside the enterprise, the Jabber client receives resolution from corporate DNS either pointing to Unified CM or Unified CM IM and Presence. In this case, Jabber will connect directly to the resolved collaboration application service node(s). When outside the enterprise, Jabber does not receive resolution for Unified CM or Unified CM IM and Presence from public DNS, but instead receives resolution for Expressway directing the client to connect to the enterprise through Expressway.

Figure 21-34 Cisco Jabber: Split DNS Resolution Inside and Outside the Enterprise



Note In cases where Cisco AnyConnect VPN is used for remote enterprise connectivity, Jabber will receive DNS query resolution from corporate DNS through the VPN tunnel and will connect directly to collaboration service nodes.

When deploying Expressway mobile and remote access for Cisco Jabber mobile clients, consider the following unsupported features and functions:

- Dual-mode hand-out

  Moving an active call from the WLAN interface of the Jabber device to the cellular voice interface is not supported over Expressway connections.

- CAPF enrollment for endpoint authentication and media and signaling encryption

  If secure media and signaling is required on the enterprise network, the Jabber device must complete CAPF enrollment while on-premises and prior to connecting over Expressway.

- Per-user or per-device access restrictions

  There is no mechanism for restricting specific users or devices from connecting through Expressway mobile and remote access. If Expressway mobile and remote access is deployed and a user has been provisioned for Jabber on the collaboration infrastructure (Unified CM and Unified CM IM and Presence), then the user may connect through Expressway.

- Session persistency

  All calls and other collaboration application connections over Expressway mobile and remote access are cleared whenever the network path changes or is lost.

- LDAP directory access

  LDAP traffic is not enabled on Expressway mobile and remote access connections. For this reason all Jabber clients are forced to use a UDS method for corporate directory access when connecting over Expressway, even if the directory access method has been configured as CDI.

If any of the above features and functions is required for the deployment, consider using AnyConnect VPN instead of Expressway for remote secure enterprise access.

Cisco Jabber and Expressway Mobile and Remote Access with Cisco AnyConnect VPN Split-Tunnel

In some cases VPN and Expressway might need to be deployed in parallel, enabling Jabber users to connect via either VPN or Expressway. In these situations, there are two methods of use. Jabber users can rely on the Expressway mobile and remote access feature for collaboration workloads and rely on VPN for all device traffic when connectivity back to the enterprise requires workloads outside of collaboration. In these scenarios, when the Cisco AnyConnect VPN client establishes a connection back to the enterprise, either due to VPN on-demand triggering or manual launch by the user, active connections are dropped and the user must wait for the Jabber client to reconnect to provisioned collaboration services over VPN before resuming use. This results in a poor user experience.

Alternatively, AnyConnect VPN and Expressway may be used simultaneously with split-tunneling to force collaboration flows through the Expressway mobile and remote access connection and all other traffic through the VPN tunnel. This alternative method often provides a better user experience because it prevents the Jabber client from disconnecting from Expressway and reconnecting over VPN when the VPN tunnel is established.

As shown in Figure 21-35, the split-tunneling afforded by this method of deployment relies on two basic principles

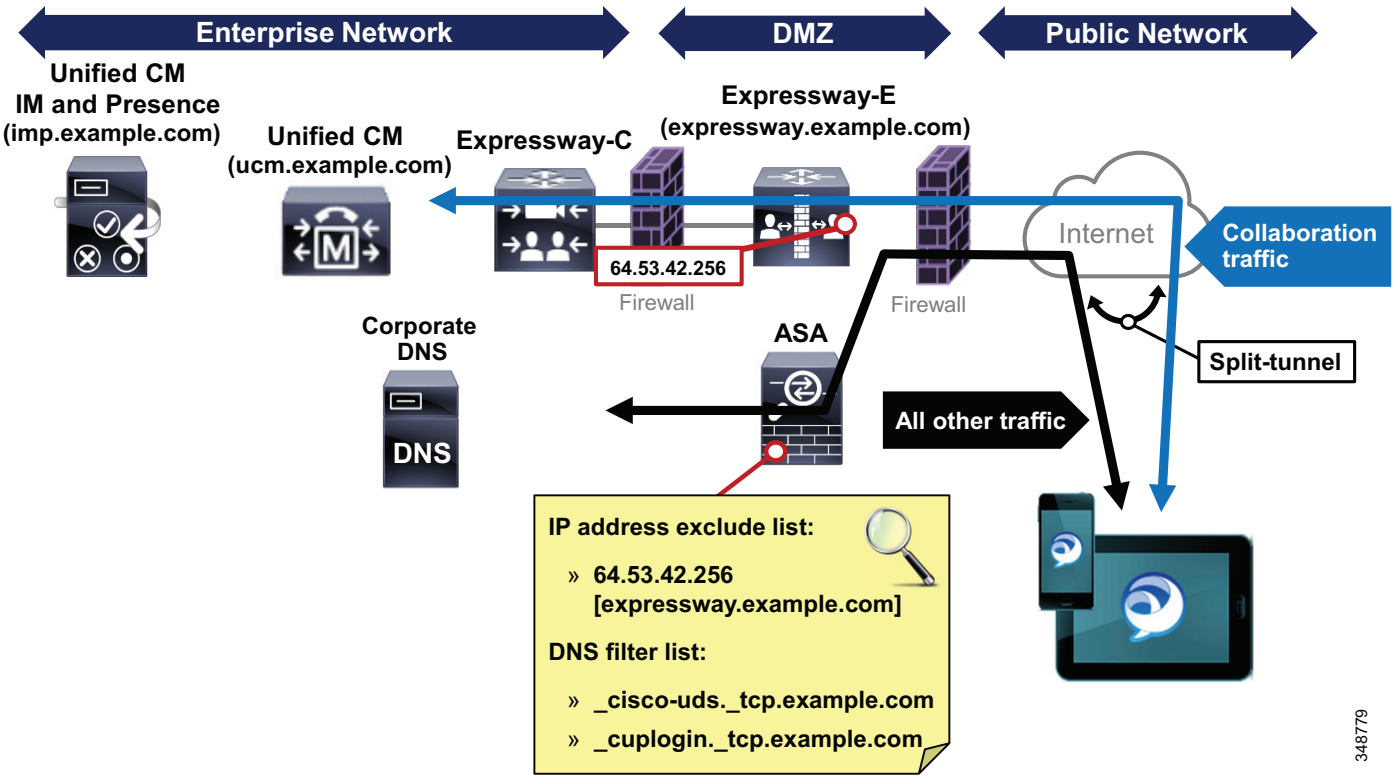
- DNS filtering at the Cisco Adaptive Security Appliance (ASA) VPN head-end

  Traffic filtering at the ASA is used to filter DNS queries from the Jabber client for _cisco-uds._tcp.<*domain*> and _cuplogin._tcp.<*domain*>. Because these DNS queries are filtered, the Jabber client is unable to resolve Unified CM or IM and Presence service record requests for direct connection to collaboration services. Therefore, the only DNS resolution will be for _collab-edge._tcp.<*domain*>, which always results in Expressway connection and traversal.

- Exclusion of Expressway access over the VPN tunnel

  IP address filtering at the ASA is used to prevent the Jabber client from connecting to the Expressway-E publicly facing interface. When filtering Expressway-E node public interface IP address(es), a split-tunnel VPN connection is created, resulting in Jabber traffic exclusion from the VPN tunnel and thus this traffic traverses Expressway while all other traffic traverses the VPN tunnel.

Figure 21-35 Cisco Jabber: Expressway Mobile and Remote Access and Cisco AnyConnect VPN



In the case of AnyConnect VPN split-tunneling with Expressway mobile and remote access, the same Expressway DNS SRV record (_collab-edge._tls) configured in the public DNS is added to the corporate DNS. This prevents the need to provide access and forward DNS queries to the public DNS through the VPN tunnel.

Although configuring an identical _collab-edge._tls SRV record in the corporate DNS would seem to violate the foundational split DNS design expected with Jabber and Expressway mobile and remote access deployments, in fact, Jabber's order of SRV resolution preference ensures appropriate behavior. Jabber's order of SRV resolution preference is for Unified CM (_cisco-uds._tcp) first, then IM and Presence (_cuplogin._tcp), and finally Expressway (_collab-edge._tls). Therefore, even when the _collab-edge._tls query can be resolved by the corporate DNS, the client will still connect directly to collaboration services because the corporate DNS will resolve queries for _cisco-uds._tcp or _cuplogin._tcp services first.

For more information about Jabber and Expressway mobile and remote access with AnyConnect VPN, refer to the information on mobile and remote access collaboration with Cisco Expressway Series, found in the *Cisco Unified Access (UA) and Bring Your Own Device (BYOD) CVD* available at

https://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Borderless_Networks/Unified_Access/BYOD_Design_Guide.html

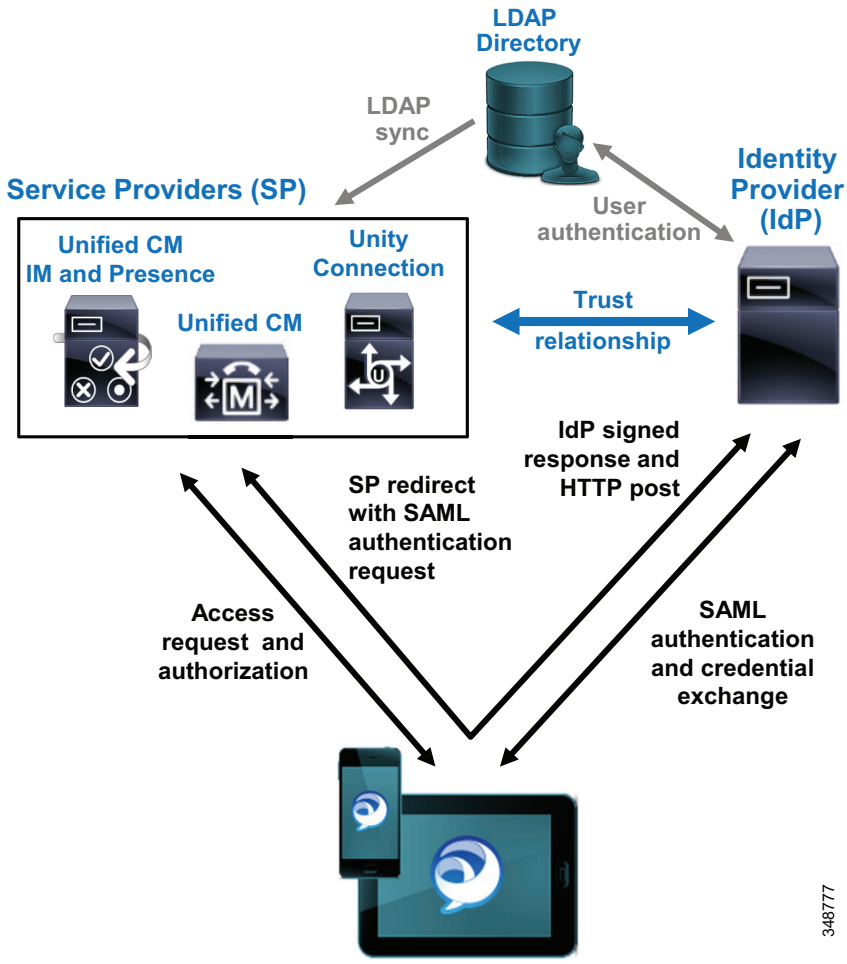
Cisco Jabber with SAML Single Sign-On

Cisco Jabber mobile clients are able to leverage single sign-on (SSO) using the Security Assertion Markup Language (SAML) version 2. Jabber and Cisco collaboration infrastructure including Unified CM, Unified CM IM and Presence, and Unity Connection leverage web-based SSO SAML v2 in order to identify and authenticate user connections, thus enabling the use of a single set of Jabber user credentials for access to all collaboration services.

As depicted in Figure 21-36, Cisco Jabber SSO depends on pre-established trust relationships between collaboration applications such as Unified CM, called service providers, and the identity provider (IdP). Unified CM and Unity Connection service providers rely on LDAP sync and integration with the corporate LDAP directory to identify users. Likewise, the IdP relies on the LDAP corporate directory for authentication of users. Supported IdPs for Cisco Jabber and collaboration services include Ping Federate, Microsoft Active Directory Federation Services (ADFS), and Open Access Manager (OpenAM).

Figure 21-36 shows a basic Jabber SSO flow. The SSO flow begins with the Jabber client requesting access to a collaboration service provider – for example, access to Unified CM for call control services. Rather than logging in directly to the collaboration service provider for access, the service provider redirects the Jabber client to the IdP with a SAML authentication request. The IdP requests authentication credentials from the Jabber user and authenticates the user against the corporate LDAP directory. Assuming that the user is authenticated successfully, the IdP returns a signed assertion which Jabber forwards to the collaboration service provider using HTTP POST. The collaboration service provider then validates the signed assertion and provides authorization to the Jabber client. For example, Jabber successfully registers to Unified CM.

Figure 21-36 Cisco Jabber with SAML SSO



In addition to forwarding a signed assertion to the Jabber client, the IdP stores a security context for the authenticated Jabber client. Should the client request access to other collaboration service providers, the IdP is able to provide subsequent signed assertions without requiring another exchange of credentials. In this way, SSO enables the Jabber user or client to access multiple collaboration services by entering their credentials once.

It is worth noting that the collaboration service provider never communicates directly with the IdP when authenticating the user.

For more information about SSO, refer to the Identity Management Architecture Overview, page 16-33, and the latest version of the *SAML SSO Deployment Guide for Cisco Unified Communications Applications,* available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

In addition to SSO user identification and authentication to on-premises collaboration applications and services, SAML SSO can also be enabled for user authentication over Expressway mobile and remote access connections. In these scenarios, an HTTPS reverse proxy is deployed in the DMZ of the enterprise to broker authentication for inbound remote access connections. The HTTPS reverse proxy

communicates with the internal enterprise IdP and brokers the SAML request and authentication exchange between the remote client and the enterprise IdP. While the HTTPS reverse proxy in the DMZ can be any generic HTTPS reverse proxy, some IdP vendors offer an option to install an IdP instance in the DMZ to serve an IdP proxy role for brokering or proxying SSO SAML requests.

Interactions Between Cisco Jabber and Cisco Unified Mobility

The Cisco Jabber mobile clients can be integrated with Cisco Unified Mobility to leverage Cisco Single Number Reach, mid-call DTMF features, two-stage dialing, and single enterprise voicemail box mobile voicemail avoidance.

Integration with Unified Mobility requires the iPhone or Android dual-mode mobile phone number to be configured within Unified CM as a mobility identity associated with the Cisco Dual Mode for iPhone or Cisco Dual Mode for Android device. Once the mobile number is configured as a mobility identity within the system, Single Number Reach can be leveraged so that incoming calls to the user's enterprise number will be extended to the iPhone or Android dual-mode device through the mobile voice network as long as the iPhone or Android dual-mode device is not connected to the enterprise and not registered to Unified CM. In situations where the dual-mode device is connected to the enterprise, registered to Unified CM, and the client calling options are set so that inbound voice-over-IP calling is enabled ("Voice over IP" or "Autoselect" when the device is connected to a WLAN), an inbound call to the enterprise number will not be extended to the mobile voice network interface of the device. When the iPhone or Android dual-mode device is connected to the enterprise, only the WLAN or mobile data interface of the device will receive the inbound call. This prevents unnecessary consumption of enterprise PSTN gateway resources.

When handling enterprise calls through the cellular voice network, the iPhone or Android dual-mode device can invoke mid-call features by means of DTMF and perform desk phone pickup for any enterprise anchored call. The dual-mode device can also leverage Mobile Voice Access and Enterprise Feature Access two-stage dialing features when making outbound calls to route these calls through the enterprise and anchor them in the enterprise PSTN gateway.

In addition to configuring a mobility identity for the iPhone or Android dual-mode device, you can configure additional mobile phone numbers or off-system phone numbers as remote destinations and associate them to the Cisco Dual Mode for iPhone or Cisco Dual Mode for Android device within Unified CM. When associating the mobility identity and additional remote destinations to the dual-mode device, you do not have to configure a remote destination profile.

When mobile users are provisioned with multiple Cisco mobile clients across multiple mobile devices (for example, a user running Cisco Jabber for Android on their Android smartphone and Cisco Jabber for iPhone and iPad on their Apple iPad), associate the mobility identity with the dual-mode device (for example, Cisco Dual Mode for Android) rather than with the tablet device (Cisco Jabber for Tablet). Because the dual-mode device leverages functionality unique to the mobility identity, including dual-mode handoff and dial via office, the mobility identity should be associated to this device. Associate all other remote destinations to the same device as the mobility identity. Associating different remote destinations on different mobile client devices for the same user makes configurations more complex and troubleshooting issues more difficult.

For more information about the Cisco Unified Mobility feature set as well as design and deployment considerations, see Cisco Unified Mobility, page 21-47.

Interactions Between Cisco Jabber and Cisco Intelligent Proximity for Mobile Voice

The Intelligent Proximity for Mobile Voice feature is designed to enable hands-free audio for the cellular or mobile line of a dual-mode devices. For this reason, usually only calls on the cellular line of the Jabber client device are enabled for hands-free audio play out on an Intelligent Proximity-capable IP endpoint. In the case of voice or video over IP calls on Cisco Jabber, Intelligent Proximity for Mobile Voice is not invoked. The one exception to this is with the Cisco IP Phone 8851 and 8861 endpoints. Because these

IP phones are audio-only, with Intelligent Proximity for Mobile Voice, audio for a Jabber IP-based call is streamed through the 8851 or 8861 phone while the video portion of this call remains on the Jabber client device. In the case of other hardware endpoints capable of Intelligent Proximity for Mobile Voice, audio for Jabber IP-based calls is not played by the IP endpoint.

Cisco Spark

The Cisco Spark mobile client is available for Android and Apple iOS mobile devices, including iPad and iPhone. Once the client application is downloaded from the appropriate application store (Apple Application Store or Google Play) and installed on the Apple iOS or Android device, users must enter their email address and activate their account with the resulting provisioning email. Once a user activates their account, the client connects to the Cisco Collaboration Cloud and the user can begin creating secure collaboration rooms with one or more people to communicate using encrypted instant messaging (IM). The user should access Cisco Spark at https://web.ciscospark.com/ using a web browser at least once in order to set a password for their account. Alternatively, the user can use the desktop Cisco Spark client available for download from https://download.ciscospark.com/. Failure to do this will require the user to activate their account via email each time they connect with the mobile client.

Cisco Spark for Android, iPad, and iPhone clients not only provide secure persistent IM collaboration rooms, but they also provide encrypted voice and video calling over IP and file sharing capabilities.

For proper Cisco Spark client operation, the mobile device must be able to reach the Internet by connecting to a wireless network (enterprise or public/private 802.11 WLAN or mobile provider data network).

Note As with Cisco Jabber, Cisco Spark mobile clients running on Apple iOS devices (iPhone and iPad) also leverage Apple Push Notification services (APNs), as described in Apple Push Notification Service (APNs) for Cisco Jabber for iPhone and iPad, page 21-99, when running in the background.

For more information about the Cisco Spark mobile clients, additional feature details, and supported hardware and software versions, refer to the Cisco Spark documentation at

https://support.ciscospark.com/

Cisco WebEx Meetings

The Cisco WebEx Meetings mobile client runs on specific Android, Apple iOS, BlackBerry, and Windows Phone mobile devices. This client enables mobile endpoints to participate in Cisco WebEx Meetings with a similar experience as with desktop browser-based Cisco WebEx Meetings. This client enables active participation in Cisco WebEx voice and video conferencing, including the ability to view participant lists and shared content.

For more information about Cisco WebEx mobile clients, refer to the product information at

https://www.cisco.com/c/en/us/products/conferencing/webex-meetings/index.html

Cisco Cloud Collaboration Services: SAML SSO for Cisco Spark and Cisco WebEx

Just as with on-premises enterprise and collaboration edge deployments described earlier, enterprise SSO can be used to facilitate secure logins to cloud collaboration services such as Cisco Spark and Cisco WebEx. With these types of deployments the enterprise IdP in combination with an HTTPS reverse proxy deployed in the enterprise DMZ leverage enterprise credentials to identify and authenticate user access to Cisco Spark and Cisco WebEx.

Cisco AnyConnect Mobile Client

The Cisco AnyConnect mobile client provides secure remote connectivity capabilities for Cisco Jabber mobile device clients, enabling connectivity over mobile data networks and non-enterprise WLANs. The Cisco AnyConnect mobile client can be downloaded from the Apple Application Store or Google Play (formerly Android Market). This client application provides SSL VPN connectivity for Apple iOS and Android mobile devices through the Cisco AnyConnect VPN solution available with the Cisco Adaptive Security Appliance (ASA) head-end.

When employing VPN network connectivity for connections over the mobile data network or public or private Wi-Fi hot spots, it is important to deploy a high-bandwidth secure VPN infrastructure that adheres to the enterprise's security requirements and policies. Careful planning is needed to ensure that the VPN infrastructure provides high bandwidth, reliable connections, and appropriate session or connection capacity based on the number of users and devices using this connectivity.

For more information on secure remote VPN connectivity using Cisco AnyConnect, refer to the Cisco AnyConnect Secure Mobile Client documentation available at

https://www.cisco.com/c/en/us/support/security/anyconnect-secure-mobility-client/tsd-products-support-series-home.html

High Availability for Cisco Mobile Clients and Devices

Although mobile devices and in particular dual-mode phones by their nature are highly available with regard to network connectivity (when the WLAN network is unavailable, the mobile voice and data networks can be used for voice and data services), enterprise WLAN and IP telephony infrastructure high availability must still be considered.

First, the enterprise WLAN must be deployed in a manner that provides redundant WLAN access. For example, APs and other WLAN infrastructure components should be deployed so that the failure of a wireless AP does not impact network connectivity for the mobile device. Likewise, WLAN management and security infrastructure must be deployed in a highly redundant fashion so that mobile devices are always able to connect securely to the network. Controller-based wireless LAN infrastructures are recommended because they enable centralized configuration and management of enterprise APs, thus allowing the WLAN to be adjusted dynamically based on network activity and AP failures.

Next, remote secure connection solution components, including the Cisco ASA head-end VPN terminator and the Cisco Expressway-E and Expressway-C nodes, should be deployed in a highly redundant fashion so that loss of a Cisco ASA or a Cisco Expressway node does not impact or prevent secure mobile and remote access connectivity for the mobile client.

Next, Unified CM call processing and registration service high availability must be considered. Just as with other devices within the enterprise that leverage Unified CM for call processing services, mobile client devices must register with Unified CM. Given the redundant nature of the Unified CM cluster architecture, which provides primary and backup call processing and device registration services, mobile device registration as well as call routing are still available even in scenarios in which a Unified CM node fails.

Similar considerations apply to PSTN access. Just as with any IP telephony deployment, multiple PSTN gateways and call routing paths should be deployed to ensure highly available access to the PSTN. This is not unique to mobile client device deployments, but is an important consideration none the less.

In the case of the Cisco Collaboration Cloud, WebEx and Cisco Spark services are highly available due to the redundant component and resource design in the cloud data centers, including both compute and network access platforms. This resilient infrastructure design provides highly reliable access for Cisco mobile clients that rely on Cisco Collaboration Cloud services.

Capacity Planning for Cisco Mobile Clients and Devices

Capacity planning considerations for Cisco mobile clients and devices, including dual-mode phones, are the same as for other IP telephony endpoints or devices that rely on the IP telephony infrastructure and applications for registration, call processing, and PSTN access services.

When deploying Cisco mobile clients and devices with Unified CM, it is important to consider the registration load on Unified CM as well as the Unified Mobility limits. A single Unified CM cluster is capable of handling a maximum of 40,000 device configurations and registrations. When deploying mobile clients and devices, you must consider the per-cluster maximum device support, and you might have to deploy additional call processing clusters to handle the added load.

In addition, as previously mentioned, the maximum number of remote destinations and mobility identities within a single Unified CM cluster is 40,000. Because most dual-mode mobile client devices will likely be integrated with Unified Mobility to take advantage of features such as Single Number Reach, single enterprise voicemail box mobile voicemail avoidance, desk phone pickup, and two-stage dialing, the mobile phone number of each of these dual-mode mobile devices must be configured as a mobility identity within the Unified CM cluster. This is necessary to facilitate integration to Unified Mobility as well as to facilitate the Handoff Number method of hand-out. Therefore, when integrating these dual-mode devices with Unified Mobility, it is important to consider the overall remote destination and mobility identity capacity of the Unified CM cluster to ensure sufficient capacity exists. If additional users or devices are already integrated to Unified Mobility within the system, they can limit the amount of remaining remote destination and mobility identity capacity available for dual-mode devices.

Another scalability consideration for Cisco mobile clients is the Cisco Expressway mobile and remote access call and proxy registration capacity of the Expressway-C and Expressway-E nodes. Expressway-C and Expressway-E clusters support a maximum of 10,000 proxy registrations and a maximum of 2,000 video or 4,000 audio calls. When determining available capacity for Cisco mobile clients, remember to include other Expressway attached devices – for example, Jabber desktop clients and fixed endpoints such as Cisco TelePresence MX and SX Series devices, and Cisco desk phones such as the 7800 and 8800 Series devices – in the calculations. Likewise, registration load on Unified CM cluster nodes must also be considered for Cisco Mobile client devices connecting to the enterprise through Expressway mobile and remote access. See Cisco Expressway, page 25-37, for more details on Cisco Expressway mobile and remote access sizing.

Overall call processing capacity of the Unified CM system and PSTN gateway capacity must also be considered when deploying mobile client devices. Beyond handling the actual mobile device configuration and registration, these system must also have sufficient capacity to handle the added BHCA impact of these mobile devices and users. Likewise, it is critical to ensure sufficient PSTN gateway capacity is available to accommodate mobile devices. This is especially the case for dual-mode mobile devices that are integrated to Unified Mobility because the types of users that would have dual-mode devices are typically highly mobile. Highly mobile users typically generate more enterprise PSTN gateway load from mobility features such as Single Number Reach, where an incoming call to a mobile user's enterprise number generates one or more calls to the PSTN, or from two-stage dialing, where a user makes a call through the enterprise by leveraging the enterprise PSTN gateway.

Finally, just as with enterprise mobility deployments, 802.11 WLAN call capacity must be considered when deploying Cisco mobile clients and device. As previously mentioned, a maximum of 27 VoWLAN calls or a maximum of 8 VVoWLAN calls are possible per 802.11 channel cell. This assumes no Bluetooth when devices are deployed on the 2.4 GHz band, 24 Mbps or higher data rates for VoWLAN calls, and 720p video resolution with bit rates up to 1 Mbps for VVoWLAN calls. Actual call capacity could be lower depending on the RF environment, wireless endpoint type, and WLAN infrastructure. See Capacity Planning for Campus Enterprise Mobility, page 21-9, for more details regarding 802.11 WLAN call capacity.

The above considerations are certainly not all unique to mobile clients and devices. They apply to all situations in which devices and users are added to Unified CM, resulting in additional load to the overall system.

For more information on general system sizing, capacity planning, and deployment considerations, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Design Considerations for Cisco Mobile Clients and Devices

Observe the following design recommendations when deploying Cisco mobile clients and devices:

- Dual-mode mobile devices must be capable of dual transfer mode (DTM) in order to be connected simultaneously to both the mobile voice and data network and the WLAN network so that the device is reachable and able to make and receive calls on both the cellular radio and WLAN interface of the device. In some cases, proper dual-mode client operation might not be possible if mobile voice and data networks do not support dual-connected devices.
- WLAN APs should be deployed with a minimum cell overlap of 20%. This overlap ensures that a mobile device can successfully roam from one AP to the next as the device moves around within a location, while still maintaining voice and data network connectivity.
- WLAN APs should be deployed with cell power level boundaries (or channel cell radius) of -67 dBm in order to minimize packet loss. Furthermore, the same-channel cell boundary separation should be approximately 19 dBm. A same-channel cell separation of 19 dBm is critical for ensuring that APs or clients do not cause co-channel interference to other devices associated to the same channel, which would likely result in poor voice and video quality.
- Whenever possible rely on the 5 GHz WLAN band (802.11a/n/ac) for connecting mobile clients and devices capable of generating voice and video traffic. 5 GHz WLANs provide better throughput and less interference for voice and video calls.
- The enterprise wired and wireless LAN should be deployed and configured to support the necessary end-to-end QoS classes of service, including priority queuing for voice media and dedicated video and signaling bandwidth, to ensure the quality of client application voice and video calls and the appropriate behavior of all features. While most clients mark traffic appropriately at Layer 3 based on Cisco QoS recommendations, appropriate Layer 2 WLAN UP marking is dependent on the client device and vendor implementation. For this reason, Layer 2 marking is not consistent across platforms and as such cannot be relied upon.
- Because mobile devices are similar to desktop computers and can generate a large variety of data and real-time traffic, these devices are typically considered untrusted. For this reason, the network should be configured to re-mark all traffic from these client devices based on port number and/or protocol. Likewise, rate limiting and policing on ingress to the network is recommended.
- Cisco recommends using only an enterprise-class voice and video optimized WLAN network for connecting mobile devices and clients. While most mobile client devices are capable of attaching to public or private WLAN access points or hot spots for connecting back to the enterprise through the Internet for call control and other collaboration services, Cisco cannot guarantee voice and video quality for these types of connections.
- When deploying Cisco collaboration mobile clients and devices on a Cisco Bring Your Own Device (BYOD) infrastructure, administrators should consider a network attachment method that does not require user intervention and which maximizes utilization of the IP telephony infrastructure. Further, for remote connectivity scenarios, all relevant ports must be opened in the corporate firewall in order for Cisco mobile clients and devices to be able to access collaboration services.

- If corporate policy dictates that the BYOD infrastructure must remotely wipe or factory-reset lost or stolen mobile devices, employees using personal mobile devices should be aware of the policy and should regularly back up personal data.
- The Unified Mobility Single Number Reach feature will not extend incoming calls to the dual-mode device's configured mobility identity if the dual-mode device is inside the enterprise and registered to Unified CM. This is by design in order to reduce utilization of enterprise PSTN resources. Because the dual-mode device registers to Unified CM, the system knows whether the device is reachable inside the enterprise; and if it is, there is no reason to extend the call to the PSTN in order to ring the dual-mode device's cellular voice radio. Only when the dual-mode device is unregistered will Single Number Reach extend incoming calls to the user's enterprise number out to the mobility identity number on the PSTN.
- When you deploy mobile devices, Cisco recommends normalizing required dialing strings so that users are able to maintain their dialing habits, whether the mobile device is connected to the enterprise or not. Because dialing on the mobile network is typically done using full E.164 (with or without a preceding '+') and mobile phone contacts are typically stored with full E.164 numbers, Cisco recommends configuring the enterprise dial plan to accommodate full E.164 or full E.164 with preceding '+' for mobile client devices. By configuring the enterprise dial plan in this manner, you can provide the best possible end-user dialing experience so that users do not have to be aware of whether the device is registered to Unified CM.
- Cisco recommends that dual-mode phone users rely exclusively on the mobile voice network for making emergency calls and determining device and user location. This is because mobile provider networks typically provide much more reliable location indication than WLAN networks. To ensure that dual-mode phones rely exclusively on the mobile voice network for emergency and location services, configure the Emergency Numbers field of the dual-mode devices within Unified CM with emergency numbers such 911, 999, and 112 in order to force these calls over the mobile voice network. Dual-mode phone users should be advised to make all emergency calls over the mobile voice network rather than the enterprise network. Although making emergency calls over corporate WLANs or mobile data networks is not recommended, mobile devices that do not have cellular voice radios are capable of making calls only through these data interfaces. Mobile devices that do not have cellular voice radios should not be relied upon for making emergency calls.
- When deploying Cisco Jabber on mobile devices, configure the WLAN network to accommodate the following deployment guidelines:
  - Minimize roaming of Cisco Jabber mobile client devices at Layer 3 on the WLAN. Layer 3 roaming, where a device IP address changes, will result in longer roam times and dropped voice packets and could even result in dropped calls.
  - Configure the same SSID across all APs utilized by the Cisco Jabber mobile client devices within the WLAN to ensure the fastest AP-to-AP roaming.
  - Configure all enterprise WLAN APs to broadcast their SSIDs in order to prevent mid-call prompts to join other APs within the WLAN infrastructure, which could result in interrupted calls.
- Provide sufficient wireless voice and video call capacity on the enterprise wireless network for Cisco mobile clients and devices by deploying the appropriate number of wireless APs to handle the desired call capacity based on mobility-enabled user BHCA rates. Each 802.11g/n (2.4 GHz) or 802.11a/n/ac (5 GHz) channel cell can support a maximum of 27 simultaneous voice-only calls with 24 Mbps or higher data rates. Each 802.11g/n (2.4 GHz) or 802.11a/n/ac (5 GHz) channel cell can support a maximum of 8 simultaneous video calls assuming 720p video resolution at up to 1 Mbps bit rate. For 2.4 GHz WLAN deployments, Bluetooth must be disabled to achieve this capacity. Actual call capacity could be lower depending on the RF environment, wireless endpoint type, and WLAN infrastructure.

- When deploying Dial via Office Reverse (DVO-R), use of the User Control method of voicemail avoidance ensures that called users do not end up in the calling user's voicemail box. This method of voicemail avoidance requires the calling user to press a number on the mobile device key pad in order to connect the DVO-R call. Failure to press a key on the mobile device results in the DVO call being cleared.
- DVO-R calls using the alternate callback number are not anchored in the enterprise and therefore desk phone pickup and DTMF-based mid-call features may not be used on these calls. In addition, voicemail avoidance is not engaged for calls to alternate callback numbers.
- The following features and capabilities are not supported over Expressway mobile and remote access connections: WLAN to cellular dual-mode handoff, LDAP directory access, per-user or per-device access restrictions, and session persistence during network path changes. If any of these features are required, consider implementing a Cisco AnyConnect VPN solution for Jabber mobile clients.
- When mobile users are provisioned with multiple Cisco mobile clients across multiple mobile devices, the mobility identity and any additional remote destinations should always be associated to the Cisco Jabber dual-mode device type.
- After initially downloading, installing, and activating the Cisco Spark account via the mobile device, the user should access Cisco Spark using a web browser or desktop client in order to create a password for their account. Once this is done, the user will be able to access Cisco Spark using any client (mobile, desktop, or web browser). Failure to set a password results in the user having to re-activate their account through email after sign-out each time.

Cisco Unified Contact Center

Revised: March 1, 2018

This chapter describes the Cisco Unified Contact Center solutions available with the Cisco Unified Communications System. It includes information on Cisco products such as Cisco Unified Contact Center Express, Cisco Unified Contact Center Enterprise, and Cisco Unified Customer Voice Portal. It also covers the design considerations for deploying these Cisco Unified Contact Center products with Cisco Unified Communications Manager and other Unified Communications components.

This chapter covers the following topics:

- Cisco Contact Center Architecture, page 22-2
- Contact Center Deployment Models, page 22-12
- Design Considerations for Contact Center Deployments, page 22-17
- Capacity Planning for Contact Centers, page 22-21
- Video Customer Care, page 22-22
- Network Management Tools, page 22-23

This chapter starts with a high-level overview of the main Cisco Unified Contact Center Portfolio. Then it covers the various Unified Communications deployment models for contact centers. Finally, it discusses design considerations on topics such as bandwidth, latency, Cisco Unified Communications Manager integration, and sizing.

The intent of this chapter is not to provide details on each contact center product and their various components but rather to discuss the design considerations for their integration with the Cisco Unified Communications System. Detailed design guidance for each Unified Contact Center product is covered in specific design guides for the Cisco Unified Contact Center Express, Cisco Unified Contact Center Enterprise, and Cisco Unified Customer Voice Portal products. Links to the product-specific design guides are listed at

https://www.cisco.com/go/srnd

What's New in This Chapter

Table 22-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 22-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
The following products have reached end of sale (EoS) and have been removed from this document: • Cisco MediaSense • Cisco Connected Analytics for Contact Center • Cisco Prime Collaboration Contact Center Assurance	For information on these products, refer to previous versions of the SRND, available at https://www.cisco.com/go/srnd.	March 1, 2018

Cisco Contact Center Architecture

This chapter discusses the following main Cisco Contact Center products and related features:

- Cisco Unified Communications Manager (Unified CM) call queuing feature
- Cisco Unified Contact Center Enterprise (Unified CCE)
- Cisco Unified Customer Voice Portal (Unified CVP)
- Cisco Unified Contact Center Express (Unified CCX)

Cisco Unified CM Call Queuing

The Cisco Unified CM call queuing feature provides the capability for queuing the incoming callers to a hunt pilot number. With this option enabled, callers to the hunt pilot can be put in queue to wait for an available agent that is configured as a hunt member to answer the call. Callers receive an initial greeting announcement when they first enter the queue, and they hear periodic announcements while they are in queue. When an agent becomes available, the call is taken out of the queue and answered by the agent. For customers who need a basic contact center with very limited functionality, Cisco Unified CM call queuing can be an option. However, unlike the full-featured Cisco Contact Center products, the Unified CM call queuing option lacks much of the contact center functionality such as agent desktop, supervisor, and reporting capabilities. If customers require complete contact center functionality, Cisco Unified Contact Center Enterprise or Cisco Unified Contact Center Express should be used.

The hunt pilot line members can display the queue status about their associated hunt pilots from the phone screen, and the queue status provides the following information:

- Hunt pilot number
- Number of calls waiting in the queue
- Longest call waiting time

In addition, Unified CM call queuing provides statistics on the number of calls currently waiting in queue and the longest call waiting time, along with other statistics, through the serviceability counters based on the hunt pilot number. This allows the supervisor to monitor the queue status using the Real Time Monitoring Tool (RTMT).

For each hunt pilot, callers can be routed to an alternate configurable destination such as voicemail or another hunt pilot if any of the following situations occurs:

- The number of calls in the queue reaches the maximum that is set by the **Maximum Number of Callers Allowed in Queue** parameter.
- The wait time of a caller in queue exceeds the threshold that is configured by the **Maximum Wait Time in Queue** parameter.
- No hunt members are logged in or registered.

Note For calls routed to the queue-enabled hunt pilot number through a SIP trunk, the SIP Rel1XX Options should be set to **Send PRACK if 1XX contains SDP** in the SIP profile associated with the SIP trunk.

For additional information on the Unified CM call queuing option, refer to the latest version of the *System Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Cisco Unified Contact Center Enterprise

Cisco Unified Contact Center Enterprise (Unified CCE) provides a contact center solution that enables you to integrate inbound and outbound voice applications with Internet applications, including real-time chat, Web collaboration, and email. This integration provides for unified capabilities, helping a single agent support multiple interactions simultaneously, regardless of the communications channel the customer has chosen. Because each interaction is unique and may require individualized service, Cisco provides contact center solutions to manage each interaction based on virtually any contact attribute. The Unified CCE deployments are typically used for large size contact centers and can support thousands of agents.

There is also a predesigned, bounded deployment model of Unified CCE: Cisco Packaged Contact Center Enterprise (Packaged CCE). Customers whose contact center requirements fit the boundaries of the solution can enjoy the advantages of the simplified management interface, smaller hardware footprint, and reduced time to install. Those customers can also benefit from the comprehensive feature set of Cisco Unified Contact Center Enterprise and Cisco Unified Customer Voice Portal. The solution comes packaged with Cisco Unified Intelligence Center for comprehensive reporting and Cisco Finesse desktop software for an enhanced, next-generation desktop experience. For more details on Packaged CCE, refer to the documentation at the following locations:

- https://www.cisco.com/en/US/products/ps12586/index.html
- https://www.cisco.com/en/US/products/ps12586/tsd_products_support_series_home.html
- https://docwiki.cisco.com/wiki/Packaged_CCE

Unified CCE employs the following major software components:

- Call Router

  The Call Router makes all the decisions on how to route a call or customer contact.

- Logger

  The Logger maintains the system database that stores contact center configurations and temporarily stores historical reporting data for distribution to the data servers. The combination of Call Router and Logger is called the *Central Controller.*

- Peripheral Gateway

  The Peripheral Gateway (PG) interfaces to various "peripheral" devices, such as Cisco Unified CM, Cisco Unified IP Interactive Voice Response (Unified IP IVR), Cisco Unified CVP, or multichannel products such as Cisco Unified Web Interaction Manager (Unified WIM) and Cisco Unified E-Mail Interaction Manager (Unified EIM). A Peripheral Gateway that interfaces with Unified CM is also referred to as an *Agent PG*.

- CTI Server and CTI Object Server (CTI OS)

  The CTI Server and CTI Object Server interface with the agent desktops. Agent desktops can be based on the Cisco Finesse agent and supervisor desktops, Finesse IP Phone Agent, or customer relationship management (CRM) connectors to third-party CRM applications.

- Administration & Data Server

  The Administration & Data Server provides a configuration interface as well as real-time and historical data storage.

The Cisco Unified CCE solution is based on the integration with Cisco Unified Communications Manager (Unified CM), which controls the agent phones. For deployments without Unified CM but with traditional ACD, use Cisco Unified Intelligent Contact Management Enterprise (Unified ICME) instead of Unified CCE.

The queuing and self-service functions are provided by Cisco Unified IP Interactive Voice Response (Unified IP IVR) or Cisco Unified Customer Voice Portal (Unified CVP) and are controlled by the Unified CCE Call Router.

Most of the Unified CCE components are required to be redundant, and these redundant instances are referred to as side A and side B instances. For example, Call Router A and Call Router B are redundant instances of the Call Router component running on two different virtual machines.

Agents can use a large variety of endpoints, including some video endpoints and some Cisco TelePresence endpoints such as the Cisco DX70 and DX80. For a list of supported endpoints, refer to the latest version of the *Unified CCE Solution Compatibility Matrix*, available at

https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-device-support-tables-list.html

Cisco Unified Customer Voice Portal

Cisco Unified Customer Voice Portal (Unified CVP) provides carrier-class voice and video IVR services on Voice over IP (VoIP) networks. It can perform basic prompt-and-collect or advanced self-service applications with CRM database integration and with automated speech recognition (ASR) and text-to-speech (TTS) integration. Unified CVP also provides IP-based call switching services by routing and transferring calls between voice gateways and IP endpoints.

Unified CVP is based on the Voice Extension Markup Language (VXML), which is an industry standard markup language similar to HTML and which is used to develop IVR services that leverage the power of web development and content delivery.

The Unified CVP solution employs the following main components:

- Unified CVP Call Server

  The Unified CVP Call Server provides call control capabilities for SIP services. The Unified CVP Call Server can also integrate with the Unified CCE Call Router through the Intelligent Contact Management (ICM) service. The IVR service provides a platform to run VXML Micro applications and to create VoiceXML pages.

- Unified CVP VXML Server

  This component executes complex IVR applications by exchanging VoiceXML pages with the VoiceXML gateway's built-in voice browser. Unified CVP VXML applications are written using Cisco Unified Call Studio and are deployed to the Unified CVP VXML Server for execution. Note that there is no RTP traffic going through the Unified CVP Call Server or the Unified CVP VXML Server.

- Cisco Voice Gateway

  The Cisco Voice Gateway is the point at which a call enters or exits the Unified CVP system. The Cisco Voice Gateway could have a TDM interface to the PSTN. Alternatively, Cisco Unified Border Element could be used when the interface to the PSTN is an IP voice trunk.

- Cisco VoiceXML Gateway

  The VoiceXML Gateway hosts the Cisco IOS Voice Browser. This component interprets VoiceXML pages from either the Unified CVP Server IVR Service or the Unified CVP VXML Server. The VoiceXML Gateway can play prompts based on .wav files to the caller and can accept input from the caller through DTMF input or speech (when integrated with Automatic Speech Recognition). It then returns the results to the controlling application and waits for further instructions.

  The Cisco VoiceXML Gateway can be deployed on the same router as the Cisco Voice Gateway. This model is typically desirable in deployments with small branch offices. But the VoiceXML Gateway can also run on a separate router platform, and this model might be desirable in large or centralized deployments with multiple voice gateways.

- Video Media Server

  A video media server in a Unified CVP comprehensive deployment enables video streaming for the Video in Queue feature. Cisco TelePresence Content Server can be used as a video media server.

Unified CVP can be deployed standalone or integrated with Unified CCE to offer voice and video self-service and queuing functions. The Unified CVP solution now supports the G.711 a-law codec end-to-end.

The Basic Video Service in Unified CVP is available when Unified CVP is deployed along with Cisco Contact Center Enterprise (Unified CCE) in a comprehensive deployment model. This service allows a video caller to interact with an audio-only IVR and subsequently connect with a video agent. It supports Cisco TelePresence endpoints such as the Cisco DX70 and DX80 as customer and agent endpoints. The video agents can also conference in a second audio-only agent by dialing a direct extension from their phone.

Video in Queue (VIQ) Basic Video is an optional feature in Unified CVP, and it can be enabled to play video to callers while they wait for a video-enabled agent or expert. Cisco TelePresence Content Server enables the video streaming. The caller can subsequently connect to a video agent.

For more information on Unified CVP system design and detailed call flows, refer to the latest version of the *Cisco Unified Customer Voice Portal Design Guide*, available at

https://www.cisco.com/c/en/us/support/customer-collaboration/unified-customer-voice-portal/products-implementation-design-guides-list.html

Cisco Unified Contact Center Express

Cisco Unified Contact Center Express (Unified CCX) meets the needs of departmental, enterprise branch, or small to medium-sized companies that need easy-to-deploy, easy-to-use, highly available and sophisticated customer interaction management for up to 400 agents. It is designed to enhance the efficiency, availability, and security of customer contact interaction management by supporting a highly available virtual contact center with integrated self-service applications across multiple sites. To simplify the deployment, Unified CCX can be preloaded on a Cisco Business Edition 6000 or 7000 system.

Unified CCX integrates with Unified CM by means of JTAPI for call control. All the Unified CCX components, including the Unified CCX engine, Unified CCX database, Finesse Server, Unified CCX Outbound Dialer, Cisco Unified Intelligence Center, and Express E-mail Manager, are installed on a single virtual machine. For system redundancy, a second identical Unified CCX instances can be added to the deployment.

Unified CCX has built-in capabilities for inbound audio and video calls, silent monitoring, and Cisco Unified Intelligence Center reporting. Additional licensing and components can enhance the solution to support outbound dialing, call recording, email, chat, social network monitoring, and workforce optimization.

Unified CCX supports advanced features such as Automated Speech Recognition (ASR) and Text to Speech (TTS), HTTP, and VXML. It also supports products such as Cisco Unified Workforce Optimization to optimize performance and quality of the contact center. Agents can use a variety of video endpoints such as the Cisco Unified IP Phone 9900 Series with camera. For a list of supported endpoints, refer to the latest version of the *Unified CCX Software Compatibility Matrix*, available at

https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-express/products-device-support-tables-list.html

Cisco Unified CCX includes IP IVR functionality for prompting, collecting, and queuing during customer inbound calls.

Cisco SocialMiner

Cisco SocialMiner is a social media customer-care solution that can help you proactively respond to customers and prospects by communicating through public social media networks such as Twitter, Facebook, or other public forums or blogging sites. By providing social media monitoring, queuing, and workflow to organize customer posts on social media networks and deliver them to your social media customer care team, your company can respond to customers in real time using the same social network the customers are using. For more information, refer to the documentation available at

https://www.cisco.com/en/US/products/ps11349/index.html

Universal Queue for Third-Party Multichannel Applications

Universal Queue describes the system's ability to route requests from various media channels to any agents in a contact center.

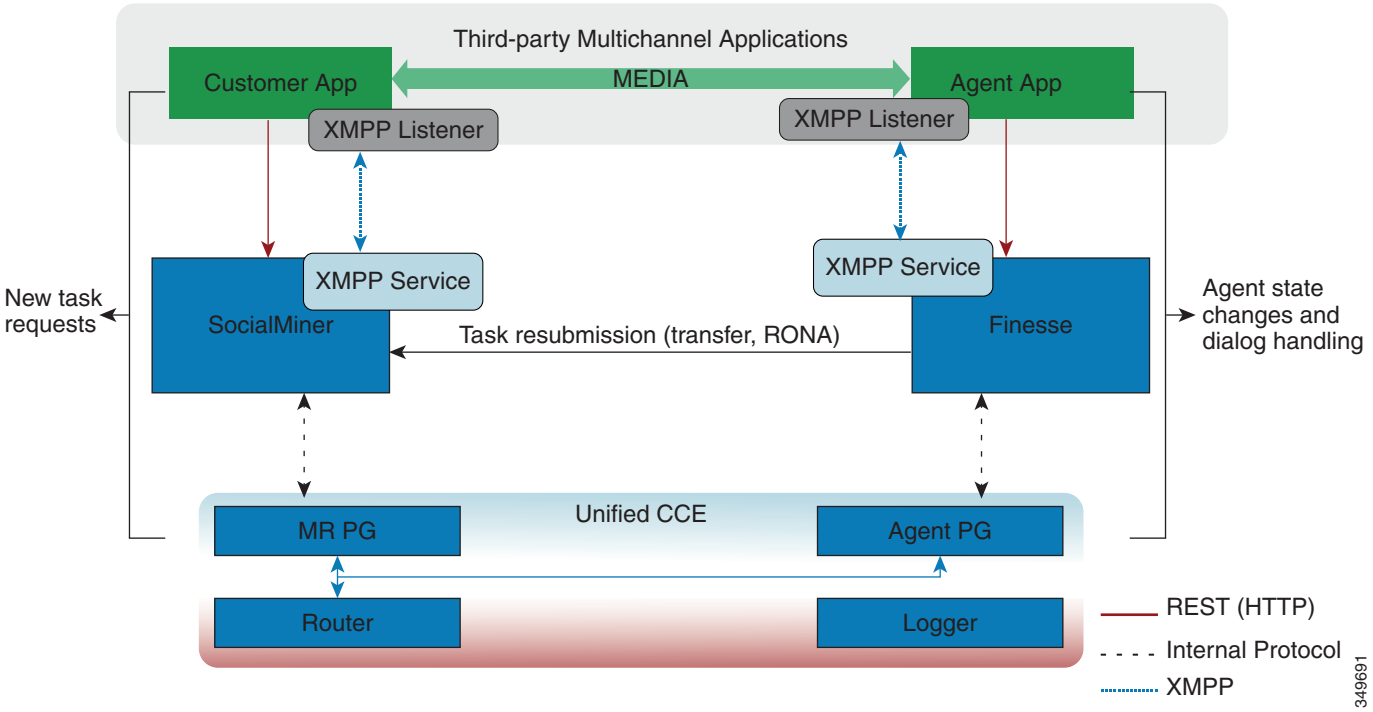
You can configure agents to handle a combination of voice calls, emails, chats, and so on. For example, you can configure an agent as a member of skill groups or precision queues in three different Media Routing Domains (MRDs) if the agent handles voice, email, and chat. You can design routing scripts to send requests to these agents based on business rules, regardless of the media. Agents signed into multiple MRDs may switch media on a task-by-task basis.

Universal Queuing APIs provide a standard way to request, queue, route, and handle third-party multichannel tasks in Unified CCE. (See Figure 22-1.)

Contact Center customers or partners can develop applications using Cisco SocialMiner and Cisco Finesse APIs in order to use Universal Queue. The SocialMiner Task API enables applications to submit non-voice task requests to Unified CCE. The Finesse APIs enable agents to sign into various types of media and handle the tasks. Agents sign into and manage their state in each media independently.

Cisco partners can use the sample code available on Cisco DevNet (https://developer.cisco.com/site/devnet/home/index.gsp) as a guide for building these applications.

Figure 22-1 Universal Queue for Third-party Multichannel Applications Solution Components



SocialMiner and Universal Queue

Third-party multichannel applications use SocialMiner's Task API to submit non-voice tasks to Unified CCE. The API works in conjunction with SocialMiner task feeds, campaigns, and notifications to pass task requests to the contact center for routing.

The Task API supports the use of Call variables and Extended Call Context (ECC) variables for task requests. Use these variables to send customer-specific information with the request, including attributes of the media such as the chat room URL or the email handle.

Unified CCE and Universal Queue

Cisco Unified CCE provides the following functionality as part of Universal Queue:

- Processes the task request
- Provides estimated wait time for the task request
- Notifies SocialMiner when an agent has been selected
- Routes the task request to an agent, using either skill group or precision queue routing
- Reports on contact center activity across media

Finesse and Universal Queue

Cisco Finesse provides Universal Queue functionality via the Media API and Dialog API. With the Media API, agents using third-party multichannel applications can:

- Sign into multiple MRDs
- Change state in multiple MRDs

With the Dialog API, agents using third-party multichannel applications can handle tasks from multiple MRDs.

Administration and Management

Cisco Contact Center products have built-in administration and management capabilities. For example, Unified CCE can be administered using either the Configuration Manager tool that is installed with Unified CCE or the web-based administration tools for simplified execution of the most common administration and management tasks in the Contact Center Enterprise environment. In addition, REST API support allows third-party developers to create applications that can control many of the administration and support tasks.

Unified CVP can be administered with the Unified CVP Operations Console, also known as Operations, Administration, Maintenance, and Provisioning (OAMP).

In addition, Cisco Unified Contact Center Management Portal (Unified CCMP) can be deployed to simplify the operations and procedures for performing basic administrative functions such as managing agents and equipment. Unified CCMP is a browser-based management application designed for use by contact center system administrators, business users, and supervisors. It is a dense multi-tenant provisioning platform that overlays the Cisco Unified CCE, Unified CM, and Unified CVP equipment.

Reporting

Cisco Unified Intelligence Center is the main reporting tool for the Cisco Contact Center solutions. It is supported by Unified CCE, Unified CCX, and Unified CVP. This platform is a web-based application offering many Web 2.0 features, high scalability, performance, and advanced features such as the ability to integrate data from other Cisco Unified Communications products or third-party data sources.

Cisco Unified Intelligence Center gets source data from a database, such as an Unified CCE Administration & Data Server database or the Unified CVP Reporting Informix database. Reports are then generated and provided to a reporting client.

Multichannel Support

The Cisco Unified Enterprise solution supports web interaction and email interaction for multichannel support. Cisco Unified Web Interaction Manager (Unified WIM) technology helps ensure that communication can be established from nearly any web browser. Cisco Unified E-Mail Interaction Manager (Unified EIM) provides inbound email routing, automated or agent assisted email responses, real-time and historical reporting, and role-based hierarchical rights management for agents, supervisors, administrators, and knowledge base administrators.

For more design information on these products, refer to the *Cisco Unified E-Mail and Web Interaction Manager Solution Reference Network Design Guide*, available at

https://www.cisco.com/en/US/products/ps7236/products_implementation_design_guides_list.html

Recording and Silent Monitoring

Cisco Unified Contact Center solutions provide recording and silent monitoring capabilities based on the following mechanisms:

- The SPAN feature in Cisco switches

  This feature replicates the network traffic to a destination port to which a Cisco contact center server is connected.

- Unified CM and media replication by the built-in-bridge (BIB) in Cisco IP Phones

  With this option, Unified CM is involved in setting up the recording flows and can perform call admission control for those flows.

- Media forking by Cisco Unified Border Element gateway

For more details on call recording and monitoring, see the chapter on Call Recording and Monitoring, page 23-1.

Contact Sharing

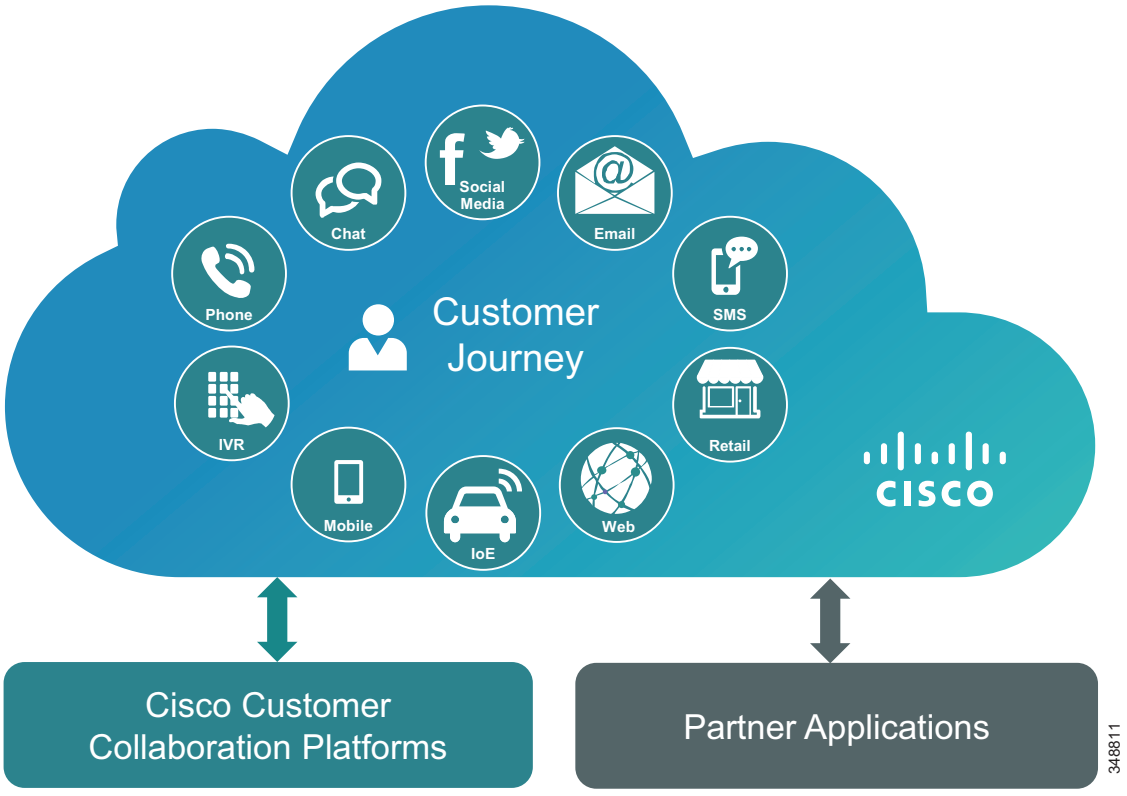
Contact Sharing enables large contact centers to grow larger. Centralized self-service (IVR ICM deployment model) uses a Contact Sharing routing node to distribute calls to two Unified CCE instances for horizontal scaling. Live Data is a prerequisite for Contact Sharing and must be installed and configured prior to use of Contact Sharing. Contact Sharing also requires the IVR Cisco Intelligent Contact Management (ICM) deployment model to be enabled in the deployment. For details about Contact Sharing, refer to the latest version of the *Cisco Unified Contact Center Enterprise Features Guide*, available at

https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-feature-guides-list.html

Context Service

Context Service is a cloud-based storage service that provides a repository for customer journey data. It enables Cisco Contact Center customers to deliver a seamless omnichannel experience through integration with other Cisco Customer Collaboration products as well as APIs for third-party integration, as depicted in Figure 22-2.

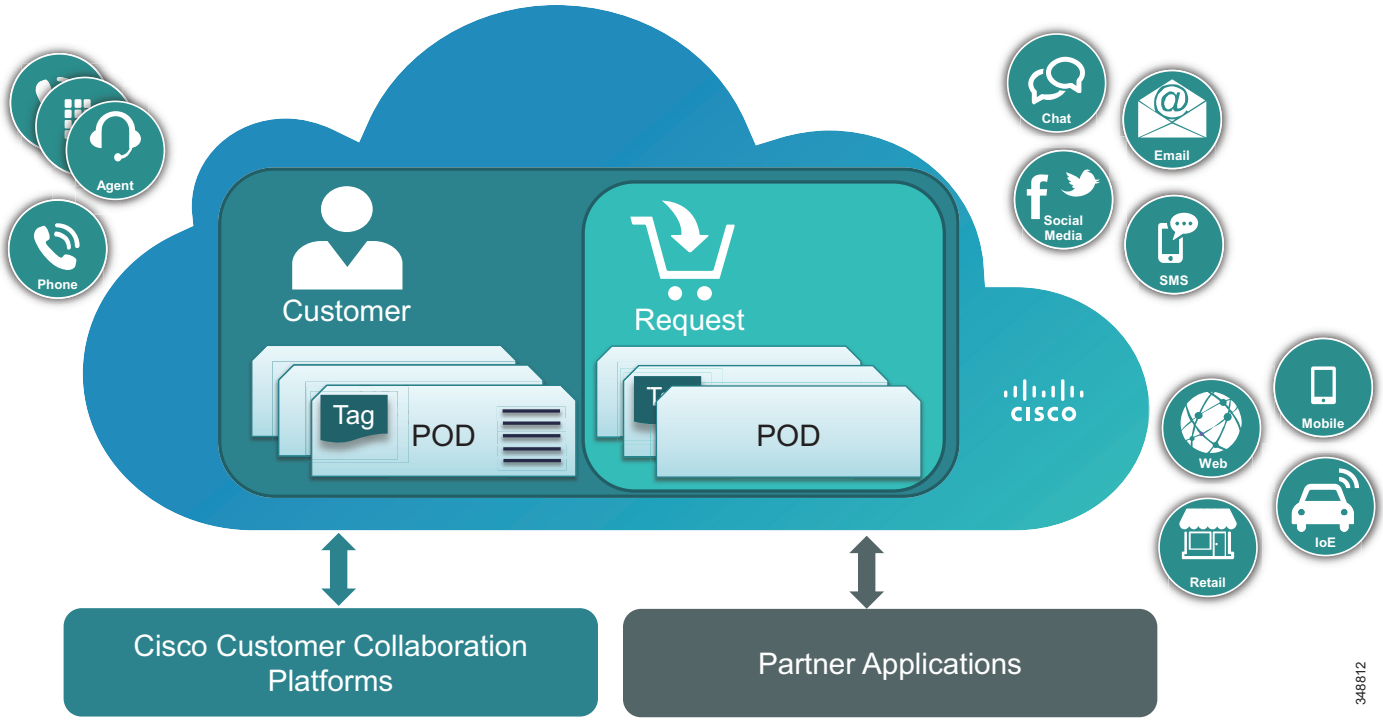
Figure 22-2 Context Service Integration



Context Service allows any application to write and read customer journey activity. Cisco Contact Center customers, referred to in this section as the *business*, have access to Context Service from within their Cisco Contact Center platforms. The Cisco Contact Center platforms are enabled and optionally configured to post context data about contact center interactions.

Context Service stores this data in an element called a Piece of Data (POD). A POD can store any metadata about the consumer interaction, except for the media (such as audio recording). Businesses choose which fields (metadata) to store in the POD and the level of data privacy for each field. PODs can be organized by customer and also grouped together as part of a collection of interactions called a Request (see Figure 22-3). Context Service also provides tagging capability to group PODs for correlation, trending, and analytics.

Figure 22-3 Pieces of Data (PODs) and Requests



Context Service is hosted on Cisco Intercloud, which is an ecosystem of Cisco and partner data centers that is managed and operated by the Cisco data center team across the globe. Context Service follows a data privacy model very similar to Cisco Spark, in which each business controls access to its data. The data is encrypted/decrypted on-premises at the client and stored as an encrypted blob in Cisco data centers. Businesses can choose to host the encryption keys (Keystore) on their premises. This is analogous to valuables stored in a safety deposit box (locker) at a bank; even though the valuables are in the bank, the customer has the key to the deposit box and controls access to it. This is a newer approach to data privacy, and it puts the customer in control without the overhead of hosting a private cloud. Context Service provides a level of data privacy classification so that businesses can store their customers' Personally Identifiable Information (PII) separately from other encrypted data, and thereby allowing businesses to provide third-party analytics vendors with controlled access to their encrypted data without giving access to their customers' PII data.

Context Service can store data for Universal Queue task contacts. When Context Service is enabled, SocialMiner selects pieces of data from an incoming task request and saves it in a POD in the cloud. You can specify the media type of the POD in the task request. If you don't specify the media type, then the media type **event** is set to the POD.

Context Service is managed by Cisco Collaboration Management (Atlas), which is the management portal for all Cisco cloud collaboration offerings, including Cisco Spark. Cisco partners and businesses use Collaboration Management to connect on-premises clients, manage the POD data model (fields), monitor POD usage, and so forth.

Context Service provides an open API and Java/JS SDK to make it easy for technology partners to integrate their applications with Context Service.

For more details on Cisco Context Service, refer to the Cisco Unified Contact Center design guides available at https://www.cisco.com/go/srnd.

Cisco Virtualized Voice Browser

Cisco Virtualized Voice Browser (Cisco VVB) provides a platform for interpreting VoiceXML documents. When a new call arrives at the contact center, the VVB allocates a VXML port that represents the VoIP endpoint. Cisco VVB sends HTTP requests to the Cisco Unified Customer Voice Portal (Unified CVP) VXML server. In response to the HTTP request, the Unified CVP VXML server executes the request and sends a dynamically generated VXML document. For more information about Cisco VVB, refer to the Cisco Virtualized Voice Browser design considerations and installation and configuration options documented in the latest version of the *Installing and Configuring Guide for Cisco HCS*, available at

https://www.cisco.com/c/en/us/support/unified-communications/hosted-collaboration-solution-contact-center/products-installation-guides-list.html

Contact Center Deployment Models

This section describes the various design models used for deploying Cisco Unified Contact Center solutions. For more details on these deployment models, refer to the Cisco Unified Contact Center design guides available at https://www.cisco.com/go/srnd.

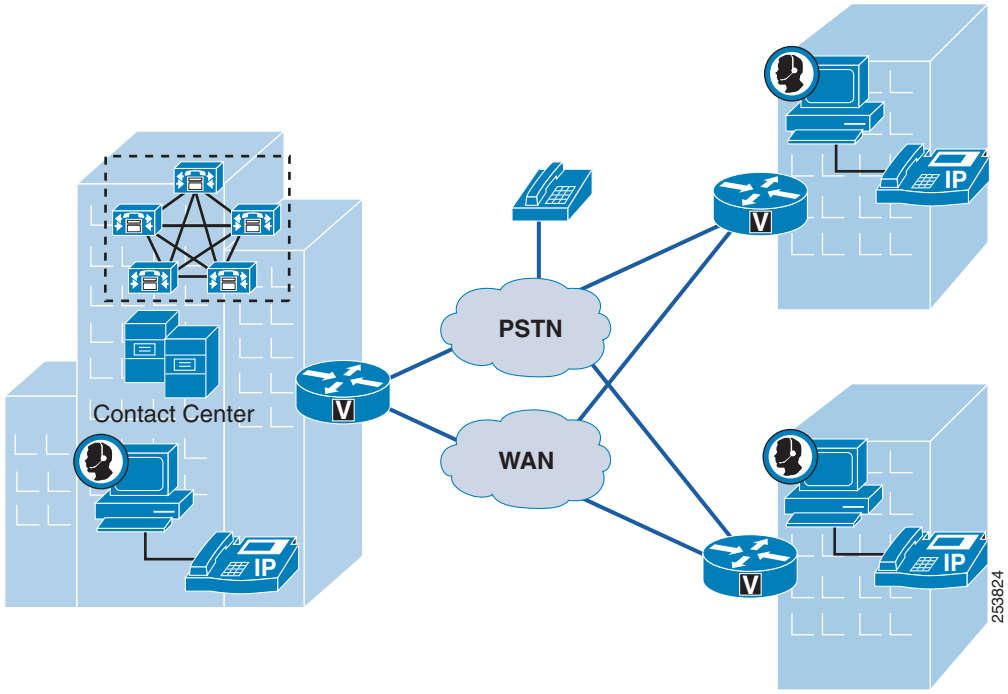
Single-Site Contact Center

In this deployment, all the components such as call processing agents, voice gateways, and contact center applications are in the same site. Agents and supervisors are also located at that site. The main benefit of the single-site deployment model is that there is no WAN connectivity required and, therefore, no need to use a low-bandwidth codec such as G.729, transcoders, compressed Real-Time Transport Protocol (cRTP), or call admission control.

Multisite Contact Center with Centralized Call Processing

A multisite deployment with centralized call processing consists of a single call processing cluster that provides services for many remote sites and uses the IP WAN. Cisco Contact Center applications (Unified CCE, Unified CCX, and Unified CVP) are also typically centralized to reduce the overall costs of management and administration. Figure 22-4 illustrates this type of deployment.

Figure 22-4 Multisite Contact Center with Centralized Call Processing



Because the agents or the voice gateways in this type of deployment are located in remote sites, it is important to consider the bandwidth requirements between the sites. It is also important to carefully configure call admission control, Quality of Service (QoS), codecs, and so forth. For more information on the general design considerations for Unified Communications solutions, refer to the chapter on Collaboration Deployment Models, page 10-1.

Contact center deployments in a Unified Communications system typically have the following additional bandwidth requirements:

- The traffic volume handled by the agents is higher than that of typical users, and therefore voice and signaling traffic is also higher for agents.
- Agents and supervisors use desktops with screen popup, reports and statistics, and so forth. This causes data traffic between the agent or supervisor desktops and the contact center servers. In addition, bandwidth calculations must account for reporting information if, for example, an agent or supervisor is remote and pulls data from a server in a central location. For more information and guidance, refer to the design guides for the individual Cisco Contact Center products, available at https://www.cisco.com/go/srnd.
- Depending on type of IVR solution, there could be traffic between the voice gateway and the IVR system. For example, if the voice gateways are distributed and calls arrive at a voice gateway located in a remote site with Unified IP IVR, there would be voice traffic across the WAN between the voice gateway and Unified IP IVR. With Unified CVP, the call could be queued at the remote site, with the VXML Gateway providing call treatment and queuing and therefore avoiding voice traffic across the WAN for IVR and reducing overall WAN bandwidth requirements.

Remote agents (for example, agents working from home) are also supported with Cisco Unified Contact Center. There are mainly two solutions. The first one requires the agent to use an IP phone that is connected to the central site by a broadband internet connection. In this solution, the phone is CTI

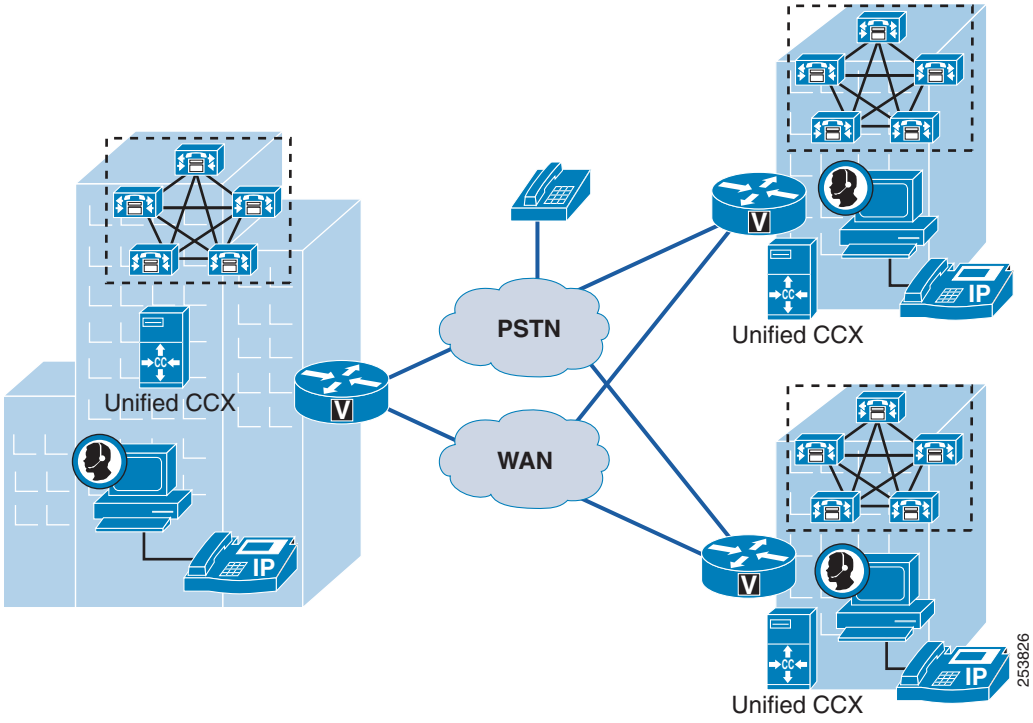
controlled by the Cisco Unified Contact Center application. The second solution is based on Cisco Unified Mobile Agent, which enables an agent to participate in a call center with any PSTN phone such as cell phone.

Multisite Contact Center with Distributed Call Processing

The model for a multisite deployment with distributed call processing consists of multiple sites, each with its own call processing cluster connected to an IP WAN. This section assumes that each Unified CM cluster has agents registered to it.

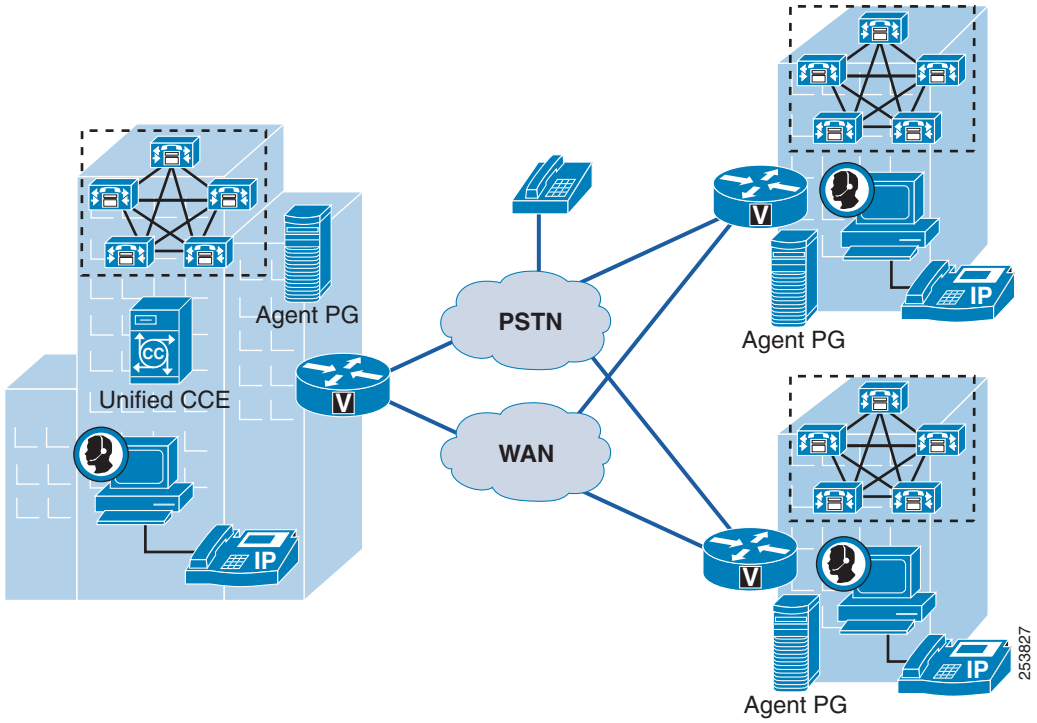
A Unified CCX deployment cannot be shared across multiple Unified CM clusters. Each Unified CM cluster requires its own Unified CCX deployment, as illustrated in Figure 22-5.

Figure 22-5 Multisite Unified CCX Deployment with Distributed Call Processing



Requirements for Unified CCE differ from Unified CCX. A single Unified CCE system can span across multiple Unified CM clusters distributed across multiple geographic locations. A Unified CCE Agent PGs must be installed in each Unified CM cluster location and could be physically remote from the Unified CCE Central Controller (Call Router + Logger). Figure 22-6 illustrates this type of deployment and highlights the placement of the Agent PG.

Figure 22-6 Multisite Unified CCE Deployment with Distributed Call Processing



If you require multiple contact center deployments, you could connect those deployments through Unified ICM by using the parent/child deployment model to form a single virtual contact center. The parent/child model provides several benefits, such as enterprise queuing and enterprise reporting across all the contact center deployments. It also provides complete site redundancy and higher scalability. For more details on the parent/child model, refer to the following documents:

- *Cisco Unified Contact Center Enterprise Design Guide*, available at

  https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-implementation-design-guides-list.html

- *Cisco Contact Center Gateway Deployment Guide for Cisco Unified ICME/CCE*, available at

  https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-installation-guides-list.html

Similarly to the multisite model with centralized call processing, multisite deployments with distributed call processing require careful configuration of QoS, call admission control, codecs, and so forth.

Clustering Over the IP WAN

In this deployment model, a single Unified CM cluster is deployed across multiple sites that are connected by an IP WAN with QoS features enabled. Cisco Unified Contact Center solutions can be deployed with this model. In fact, the Cisco Unified Contact Center components themselves can also be clustered over the WAN.

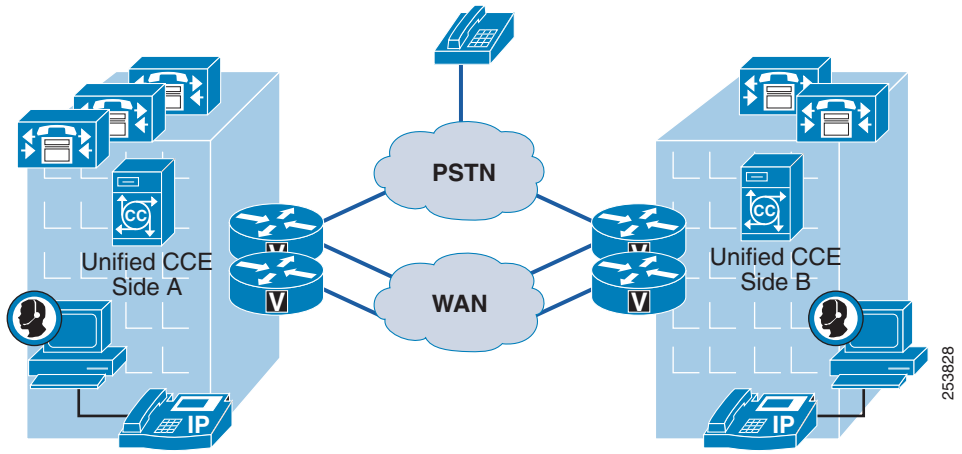
For example, with Unified CCE, the side A components could be remote from the Unified CCE side B components and separated from them by an IP WAN connection. (For more details on Unified CCE high availability, see High Availability for Contact Centers, page 22-17.) The following design considerations apply to this type of deployment:

- The IP WAN between the two sites must be highly available, with no single point of failure. For example, the IP WAN links, routers, and switches must be redundant. WAN link redundancy could be achieved with multiple WAN links or with a SONET ring, which is highly resilient and has built-in redundancy.
- The Agent Peripheral Gateway (PG) and the CTI Manager to which it is connected must be located in the same data center. Because of the large amount of redirect and transfer traffic and additional CTI traffic, the Intra-Cluster Communication Signaling (ICCS) bandwidth requirements between the Unified CM nodes are higher when deploying Unified CCE.
- If the primary Unified CCE and Unified CM nodes are located in one site and the secondary Unified CCE and Unified CM nodes are in another site, the maximum latency between the two sites is dictated by the Unified CM latency requirement of 80 ms round trip time (RTT). However, if the Unified CCE nodes are in different locations than the Unified CM nodes, it is possible to have a higher latency between the redundant Unified CCE nodes.

Figure 22-7 illustrates a deployment of Unified CCE using clustering over the WAN. For more details, refer to the *Cisco Unified Contact Center Enterprise Design Guide*, available at

https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-implementation-design-guides-list.html

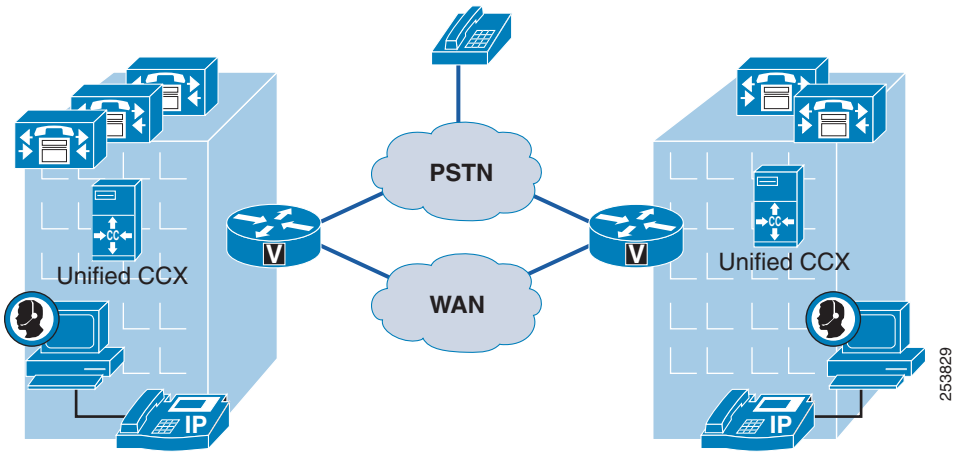
Figure 22-7 Unified CCE Deployment with Clustering Over the WAN



With Unified CCX and Unified IP IVR solutions, the primary Unified CCX or Unified IP IVR node could also be remote from the backup node. The requirements for Unified CCX deployments are different than the ones for Unified CCE deployments. For example, redundant WAN links are not required with Unified CCX. Also, the maximum latency between the primary and backup Unified CCX nodes is 80 ms RTT. Figure 22-8 illustrates this type of deployment. For more details, refer to the *Cisco Unified Contact Center Express Design Guide*, available at

https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-express/products-implementation-design-guides-list.html

Figure 22-8 Unified CCX Deployment with Clustering Over the WAN



Design Considerations for Contact Center Deployments

This section summarizes the following major design considerations for contact center deployments:

- High Availability for Contact Centers, page 22-17
- Bandwidth, Latency, and QoS Considerations, page 22-18
- Call Admission Control, page 22-19
- Integration with Unified CM, page 22-20
- Other Design Considerations for Contact Centers, page 22-20

High Availability for Contact Centers

All Cisco Unified Contact Center products provide high availability. For example, with Unified CCX or Unified IP IVR, you could add a second identical Unified CCX or Unified IP IVR node to provide high availability. The second node can reside in the same data center as the primary node or, if geographic redundancy is required, the second node can reside in a different data center across the WAN from the primary node (see Clustering Over the IP WAN, page 22-15). One of the nodes would be the active node and would handle all the call processing. The other node would be in standby mode and become active only if the primary node fails. Unified CVP also supports high available deployments with multiple Unified CVP nodes, voice gateways, VXML gateways, SIP proxies, and so forth.

With Unified CCE, most of the components are required to be redundant, and the redundant instances are referred to as side A and side B instances. For example, Call Router A and Call Router B are redundant instances of the Call Router module (process) running on two different virtual machines. This redundant configuration is also referred to as *duplex mode*. The Call Routers run in synchronized execution across the two instances, which means both sides of the duplex instances process every call. Other components, such as the Peripheral Gateways, run in hot-standby mode, meaning that only one of the Peripheral Gateways is actually active at any given time.

In addition to the redundancy of the Unified Contact Center components themselves, their integration with Unified CM can also be redundant. For example, each Unified CCX or Unified IP IVR node can connect to a primary CTI Manager and also to a backup CTI Manager in case the primary CTI Manager fails. With Unified CCE, a PG side A would connect to a primary CTI Manager, while the redundant PG side B connects to the secondary CTI Manager, thus providing high availability if one CTI Manager fails.

For more details, refer to the Cisco Unified Contact Center design guides available at https://www.cisco.com/go/srnd.

Bandwidth, Latency, and QoS Considerations

This section describes how to provision WAN bandwidth in a multisite contact center deployment, taking into account various types of call control traffic and real-time voice traffic. It is important to understand the latency and QoS parameters because adequate bandwidth provisioning and implementation of QoS are critical components in the success of contact center deployments.

Bandwidth Provisioning

Contact center solutions require sufficient WAN bandwidth to accommodate the following main types of traffic:

- Voice traffic between the ingress gateway and the IVR system. With Unified IP IVR, if the Unified IP IVR cluster is in a central location and PSTN gateways are in remote locations, there will be voice traffic over the WAN. With Unified CVP, it is possible to queue the call at the edge and therefore keep the voice traffic local to the remote site to avoid voice traffic across a WAN link. Video queuing is also supported with the Unified CVP Video in Queue (ViQ) feature, so also consider the video traffic between the caller and the video media server.
- Voice traffic between the ingress gateway and the agent, or voice traffic between the caller and agent for internal calls. There could also be video traffic between the caller and the agent if the contact center deployment supports video.
- Voice and video signaling traffic. This is typically for the signaling traffic between the ingress gateway or caller endpoint and Unified CM, and between the agent phone and Unified CM.
- VXML Gateway traffic if Unified CVP is deployed. The traffic includes media file retrieval from the media server and VXML documents exchanged with the VXML server.
- Data traffic between the Finesse agent or supervisor desktop and the application server(s) hosting Finesse gadgets.
- Reporting traffic between the reporting user and the Unified Contact Center Reporting server.
- Traffic between Unified Contact Center servers if they are remote from each other. For example, this type of traffic occurs with clustering over the IP WAN or with multisite and distributed call processing with PGs remote from the Unified CCE Central Controller.
- Additional Intra-Cluster Communication Signaling (ICCS) traffic between the Unified CM subscribers due to the large amount of redirect and transfer traffic and additional CTI traffic.
- Voice traffic due to recording and silent monitoring. Depending on the solution, one or two RTP streams could be sent in order to silently monitor or record the conversation with an agent.

Bandwidth calculations and guidelines are provided in the Cisco Unified Contact Center design guides available at https://www.cisco.com/go/srnd.

Latency

Agents and supervisors can be located remotely from the call processing components and the contact center. Technically, the delay between the Finesse Server and the Finesse desktop could be very high because of high time-out values. Long latency will affect the user experience and might cause confusion or become unacceptable from the user perspective. For example, the phone could start ringing but the desktop might not be updated until later.

Latency requirements between the contact center and the call processing components, and between the contact center components themselves, depend on the contact center solutions. For example, the Unified CCX redundant nodes can be located remotely from each other, with a maximum latency of 80 ms RTT. With Unified CCE, the maximum latency between the Unified CCE components and Unified CM, or between the Unified CCE components themselves, is greater than 80 ms RTT.

For more details, refer to the Cisco Unified Contact Center design guides available at https://www.cisco.com/go/srnd.

QoS

Similar to deployments with other Unified Communications components, contact center deployments require the configuration of Quality of Service (QoS) to prioritize time-sensitive or critical traffic. QoS marking for voice and voice signaling in a contact center environment is the same as with other Unified Communications deployments. Traffic specific to the contact center must be marked with specific QoS markings. For example, some of the traffic for the Unified CCE private network must be marked as AF31, while other traffic must be marked as AF11. The QoS marking recommendations and QoS design guidance are documented for each Unified Contact Center solution in their respective Cisco Unified Contact Center design guides available at https://www.cisco.com/go/srnd.

Call Admission Control

Similar to deployments with other Unified Communications components, contact center deployments require careful provisioning of call admission control. The same mechanisms described in the chapter on Bandwidth Management, page 13-1, also apply to contact center environments.

Voice traffic associated with silent monitoring and recording might not be accounted for in the call admission control calculation. For example, voice traffic from silent monitoring and recording by Unified CM (voice traffic forked at the phone) is properly accounted for, but voice traffic from desktop-based silent monitoring (desktop connected to the back of the agent IP phone) is not counted in call admission control calculations.

Call admission control for Mobile Agent and Unified CVP involves special considerations. For more details, refer to the Cisco Unified Contact Center design guides available at https://www.cisco.com/go/srnd.

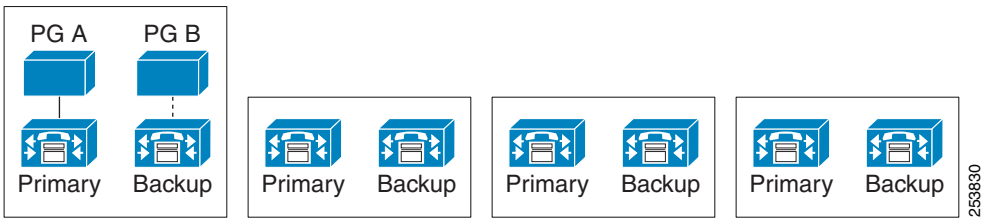
Integration with Unified CM

The following design considerations apply when integrating Cisco Unified Contact Center components with Unified CM:

- For administration and upgrade purposes, Cisco recommends separate Unified CM clusters for contact center and non-contact center deployments. If separate clusters are not possible, then Cisco recommends separate Unified CM subscriber nodes for contact center and non-contact center applications.
- With contact center deployments, Cisco recommends that you do not use a 2:1 redundancy scheme for Unified CM. Use 1:1 redundancy to provide higher resiliency and faster upgrades.
- The integration between Unified CM and Unified CCX, Unified IP IVR, or Unified CCE is done through JTAPI. The Unified CCX cluster connects to a primary CTI Manager and also has a backup connection to a secondary CTI Manager. With Unified CCE, the Agent PG connects to only one CTI Manager. The redundant Agent PG connects to the backup CTI Manager only. If the primary CTI Manager fails, the primary Agent PG will also fail and trigger the failover.
- A single PG can control and monitor agent phones on all Unified CM subscriber pairs in a centralized deployment, as illustrated in Figure 22-9.
- It is possible to integrate multiple Unified CCX deployments with a single Unified CM cluster.

For more details on Unified CM integration, refer to the Cisco Unified Contact Center design guides available at https://www.cisco.com/go/srnd.

Figure 22-9 Deployment with One Agent PG and Four Unified CM Subscriber Pairs



Other Design Considerations for Contact Centers

The following additional design considerations apply in the situations indicated:

- Because Unified CVP allows queuing at the edge, deploying Unified CVP instead of Unified IP IVR could lower the bandwidth requirements for multisite deployments.
- Most of the Cisco Unified Contact Center products and components can be installed in a virtualized environment based on VMware.
- Media termination point (MTP) resources might be required in some scenarios. For example, with Mobile Agents and inbound calls through SIP trunks, MTPs are required for the associated CTI ports when RFC 2833 is negotiated. MTPs are also required in some scenarios with Unified CVP. With Unified CCX Extend and Connect, MTPs are required for the associated CTI Remote Device when RFC 2833 is negotiated.

- Transcoders might be required. For example, if phones in a WAN- connected location support only the G.729 codec but Unified CVP is configured for G.711 support, then Unified CM will engage transcoders. However, an inbound call that arrives from a gateway or Cisco Unified Border Element can start with G.711 at Unified CVP then later renegotiate to G.729 with the agents without the need for transcoders.
- Some third-party contact center products are also supported with Unified CM. The integration with Unified CM could be based on JTAPI and could use CTI ports for call treatment and queuing and CTI route points. To size Unified CM correctly, it is important to have a good understanding of the call flows and their impact on Unified CM. It is also important to understand how the redundancy is implemented and whether or not it impacts Unified CM or CTI scalability.

For more detailed design considerations, refer to the Cisco Unified Contact Center design guides available at https://www.cisco.com/go/srnd.

Capacity Planning for Contact Centers

All deployments must be sized with the Cisco Collaboration Sizing Tool. This tool performs sizing of the contact center products such as Unified CCE, Unified IP IVR, Unified CVP, and Unified CCX. It determines the contact center resources required for your deployment, such as number of agents, number of IVR ports, and number of gateway ports. In addition to performing sizing for the contact center components themselves, the tool also sizes the rest of the Unified Communications solution, including Unified CM and voice gateways. This tool is available to Cisco employees and partners only (with proper login authentication) at https://cucst.cloudapps.cisco.com/landing.

In general, sizing of the contact center depends heavily on the busy hour call attempts (BHCA) for calls coming into the contact center. It also depends on other parameters such as the Service Level Goal and Target Answer Time. For example, a deployment where 90% of the calls must be answered within 30 seconds will require more contact center resources than a deployment where 80% of the calls must be answered within 2 minutes. Another parameter that impacts the sizing is whether Finesse or Finesse IP Phone Agent is used. Use the Unified CST for sizing, and consult the respective Cisco Unified Contact Center design guides, available at https://www.cisco.com/go/srnd, for more details.

The contact center design also impacts Unified CM sizing. The following considerations apply to sizing Unified CM when it is deployed in contact center solutions:

- The maximum number of Unified CCE agents in a single Unified CM cluster depends on the IVR solution. With Unified IP IVR, CTI route points and CTI ports are used during the call treatment queuing, which consume Unified CM resources. With Unified CVP, the call treatment and queuing are typically handled by the VXML Gateway, Unified CVP VXML server, and Unified CVP call server, with no impact on Unified CM. Therefore, a single Unified CM cluster can support more agents with Unified CVP than with Unified IP IVR.
- The Unified CCE Mobile Agent feature relies on CTI ports and therefore needs additional resources from Unified CM subscribers. Therefore, Unified CM scalability is reduced when Mobile Agents are deployed.
- With Unified CCE deployments, SIP dialing is supported. With the SIP dialer, each outbound call is placed directly from the SIP dialer port to the egress voice gateway. The call reaches Unified CM only when the call is transferred to an agent. Therefore, Unified CM capacity is much higher when the SIP dialer is used.

- When sizing Unified CM, it is also important to account for any additional CTI applications. For example, some PC clients can control a phone remotely through CTI. Some call recording applications can also integrate directly with Unified CM through the CTI Manager and can monitor agent phones, which could require additional resources from Unified CM. For more details, refer to Computer Telephony Integration (CTI), page 9-28, and to the Cisco Unified Contact Center design guides available at https://www.cisco.com/go/srnd.
- Some silent monitoring and recording solutions (such as the silent monitoring and recording feature based on Unified CM) consume resources from Unified CM, whereas other solutions such as SPAN or desktop silent monitoring and recording do not.
- Again, due to the complexity associated with sizing, all deployments must be sized with the Cisco Collaboration Sizing Tool, available to Cisco employees and partners only (with proper login authentication) at https://cucst.cloudapps.cisco.com/landing.

For more details, refer to the Cisco Unified Contact Center design guides available at https://www.cisco.com/go/srnd.

Video Customer Care

For high-touch customer engagements, enhancing the customer care options to include video-enabled customer experiences can greatly improve the interaction for the both the customer and the agent.

Cisco Remote Expert Solution

The Cisco Remote Expert Solution enables customers as well as internal employees to connect with experts across multiple channels. It also delivers a consistent, interactive experience that helps optimize revenue, improve expert productivity, and build customer loyalty. Cisco Remote Expert creates a virtual pool of specialists, manages their availability, and quickly connects customers with experts across multiple channels and devices, using high-quality audio and video.

Designed to deliver a consistent customer and employee experience across multiple touch points and devices, the Cisco Remote Expert Solution is an end-to-end, multichannel collaboration platform that establishes a new industry benchmark for customer care and provides the following benefits:

- Improved response time

  Customers can reach your experts over video with the touch of a button from personal devices, from kiosks, or from customer workstations within your store, branch, or clinic.
- Increased sales close ratios

  Cisco Remote Expert can intelligently route customers to the right resource required to satisfy product and service inquiries.
- Improved cross-sell and up-sell opportunities

  Customers can engage with highly trained experts who can address the customers' needs and suggest adjacent products and services.
- Increased productivity

  Subject-matter experts can use a single platform to reach customers, regardless of device or location.

The Cisco Remote Expert Solution employs industry-leading, high-quality collaboration products and services supported by a Cisco Validated Design reference architecture and partner ecosystem.

For more details on Remote Expert, refer to the *Cisco Remote Expert Solution Design Guide*, available at

https://www.cisco.com/c/en/us/solutions/enterprise/design-zone/remote_expert.html

Network Management Tools

Unified CCE is managed with the Simple Network Management Protocol (SNMP). Unified CCE devices have a built-in SNMP agent infrastructure that supports SNMP v1, v2c, and v3, and it exposes instrumentation defined by the CISCO-CONTACT-CENTER-APPS-MIB. This MIB provides configuration, discovery, and health instrumentation that can be monitored by standard SNMP management stations. Moreover, Unified CCE provides a rich set of SNMP notifications that alert administrators of any faults in the system. Unified CCE also provides a standard syslog event feed (conforming to RFC 3164) for those administrators who want to take advantage of a more verbose set of events.

For more information about configuring the Unified CCE SNMP agent infrastructure and the syslog feed, refer to the *SNMP Guide for Cisco Unified ICM/Contact Center Enterprise & Hosted*, available at

https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-installation-and-configuration-guides-list.html

Unified CVP health monitoring can be performed by using any SNMP standard monitoring tool to get a detailed visual and tabular representation of the health of the solution network. All Unified CVP product components and most Unified CVP solution components also issue SNMP traps and statistics that can be delivered to any standard SNMP management station or monitoring tool.

Unified CCX can also be managed with SNMP and a syslog interface.

Cisco Prime Collaboration can also help manage a Contact Center deployment. For example, Cisco Prime Collaboration Assurance can be used to monitor the number of active calls, number of inbound calls per second, or number of agents logged on.

Also, the Prime Contact Center Assurance Module can be added if Prime Collaboration Assurance Advanced has previously been implemented. The Prime Contact Center Assurance Module diagrams the topology of your customer care environment and the relationship between components. It provides event correlation to speed up error root cause analysis, provides a performance dashboard to help detect and fix performance issue, and provides call trace analysis to help identify devices that break a call flow.

Call Recording and Monitoring

Revised: March 1, 2018

Call monitoring and recording solutions provide a way to monitor and record audio and video calls that traverse various components in a Unified Communications and Collaboration solution, such as Cisco IP Phones, Cisco Unified Border Element devices, or Cisco switches. These recordings can then be used by call centers and other enterprise functions for various purposes such as compliance, transcription, speech analysis, podcasting, and blogging. This chapter provides an overview of various call recording solutions available for Cisco Unified Communications and Collaboration solutions for both audio and video calls. The chapter also outlines basic design considerations for call recording solutions embedded within a Cisco Unified Communications and Collaboration solution.

What's New in This Chapter

Table 23-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 23-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Cisco MediaSense has reached end of sale (EoS) and has been removed from this chapter.	For information on Cisco MediaSense, refer to previous versions of the SRND, available at https://www.cisco.com/go/srnd.	March 1, 2018
Cisco TelePresence Content Server (TCS) has reached end of sale (EoS) and has been removed from this chapter.	For information on Cisco TCS, refer to previous versions of the SRND, available at https://www.cisco.com/go/srnd.	March 1, 2018

Types of Monitoring and Recording Solutions

This section describes the following types of call recording and monitoring solutions:

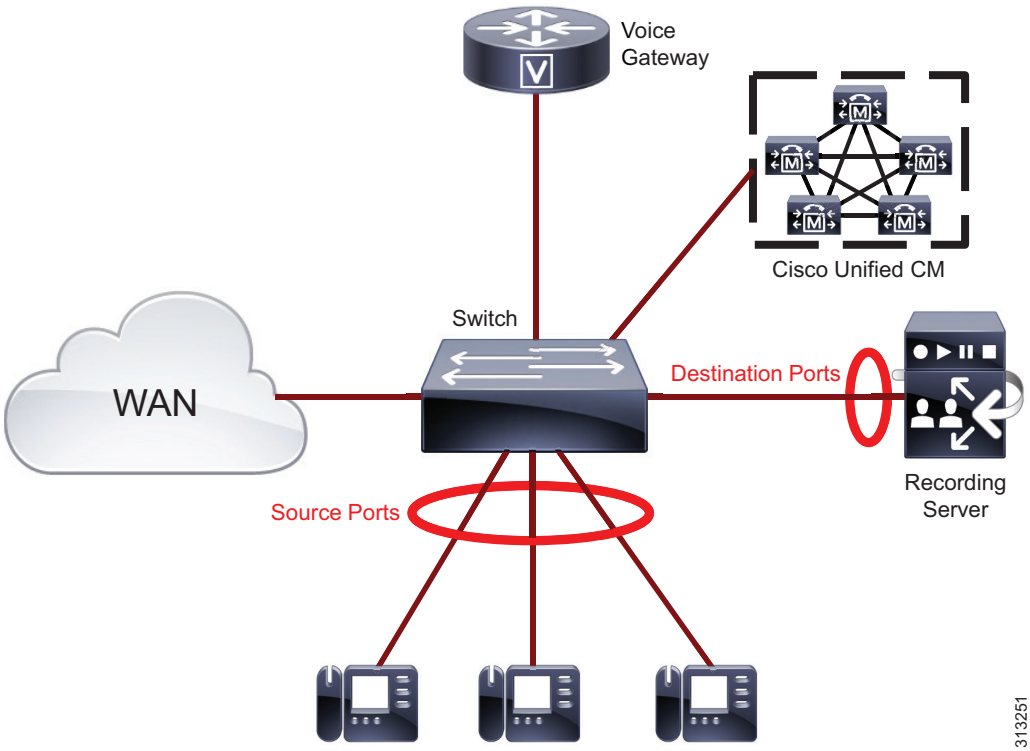
- SPAN-Based Solutions, page 23-2
- Unified CM Silent Monitoring, page 23-4
  - Unified CM Network-Based Recording, page 23-4
  - Unified CM Network-Based Recording with Built-in Bridge, page 23-6
  - Cisco Unified CM Network-Based Recording with a Gateway, page 23-7
- Agent Desktop, page 23-10

SPAN-Based Solutions

Recording solutions based on a Switched Port Analyzer (SPAN) use the packet sniffing technology for recording calls. SPAN is a method of monitoring network traffic. When SPAN is enabled on a switch port or VLAN, the switch sends a copy of all network packets traversing that port or VLAN to another port where a recording or monitoring server (such as Cisco Unified Workforce Optimization Quality Management or a third-party recording server, for example) analyzes those packets. It detects and decodes the VoIP RTP packets embedded in the network traffic and stores them as audio on a storage device. SPAN can be enabled on the ports connected to a Cisco Voice Gateway or Cisco IP Phones, as required. For example, for recording internal calls between IP phones, SPAN should be enabled on switch ports connected to the IP phones.

Figure 23-1 illustrates a SPAN-based recording solution deployment for recording internal calls. The ports marked as source ports connected to IP phones are mirrored to the destination port connected to the recording server.

Figure 23-1 SPAN-Based Recording Call Flow for Internal Calls



Several Cisco partners provide SPAN-based recording servers and applications for Cisco Unified Communications and Collaboration solutions. For technical details, refer to the specific partner product information in the *Cisco Developer Network Marketplace Solutions Catalog*, available at

https://marketplace.cisco.com/catalog/search?utf8=%E2%9C%93&x=48&y=6&search%5Btechnology_category_ids%5D=1900

In addition, network traffic flow needs to be considered for appropriate bandwidth provisioning when port mirroring is enabled.

SPAN-Based Recording and Virtualization

This section reviews some common SPAN-based deployments with virtualization enabled and lists some of the limitations. VMware provides support for the SPAN feature on VMware vSphere Distributed Switch (VDS) starting with vSphere 5.0.

In a virtualized setup, some of the Unified Communications applications, contact center applications, and the port analyzer application may be deployed on virtual machines on the same host or on different hosts. There are some limitations to SPAN-based recording solutions in a virtualized setup. For example, the following features are not supported for deployments of Cisco Unified Contact Center Enterprise (Unified CCE) with virtualization:

- Remote silent monitoring
- SPAN-based silent monitoring and recording on Cisco Unified Computing System (UCS) B-Series chassis

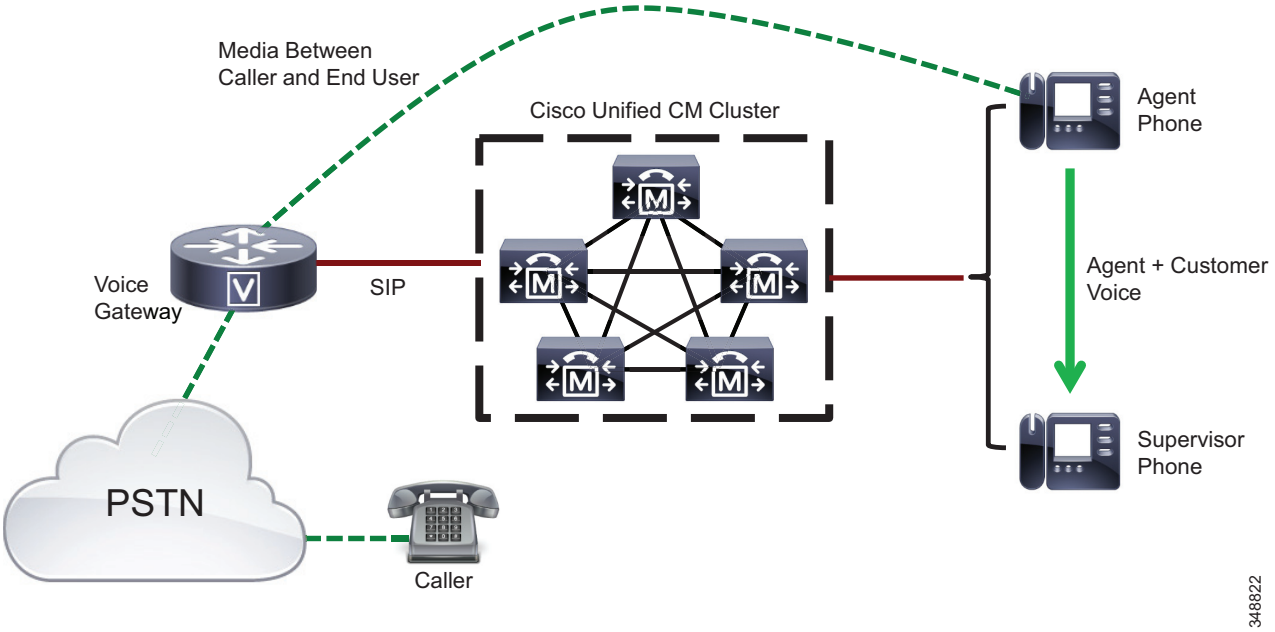
Note SPAN-based silent monitoring and recording is not supported on the UCS B-Series chassis.

Unified CM Silent Monitoring

The Unified CM Silent Monitoring feature allows a supervisor to listen to a conversation between an agent and a customer with neither the agent nor the customer aware of the supervisor's presence on the call. During call monitoring, the agent phone combines the two voice RTP streams (one for the agent and one for the customer) on the agent phone and sends the resulting stream to the supervisor phone. In addition, whisper coaching allows the supervisor to talk to the agent during the call monitoring session. Call monitoring and whisper coaching can be invoked by call center applications through the JTAPI or TAPI interfaces of Unified CM.

Figure 23-2 illustrates the basic setup for Cisco Unified CM Silent Monitoring.

Figure 23-2 Unified CM Silent Monitoring Architecture



Unified CM Network-Based Recording

The Unified CM network-based recording feature allows system administrators to record conversations between calling and called parties. Network-based recording allows for forking media using either the built-in bridge (BIB) of a supported IP phone model or a SIP gateway of a supported version and configuration. The administrator can set a preference to one forking device type or the other; however, if the preferred forking device is not available, Unified CM automatically fails over to the other method. For example, if an IP phone has recording enabled with **Phone Preferred** but there is no recording resource available (the phone does not have a built-in bridge), the gateway would be used for call recording.

Regardless of the media forking devices used by Unified CM for call recording, Unified CM always provides the metadata about the near-end and far-end parties of the recorded calls to the recording server. The metadata resides in the FROM header of the SIP Invite and other SIP messages that are sent between Unified CM and the recording server.

For details about Unified CM silent call monitoring and call recording features, refer to the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

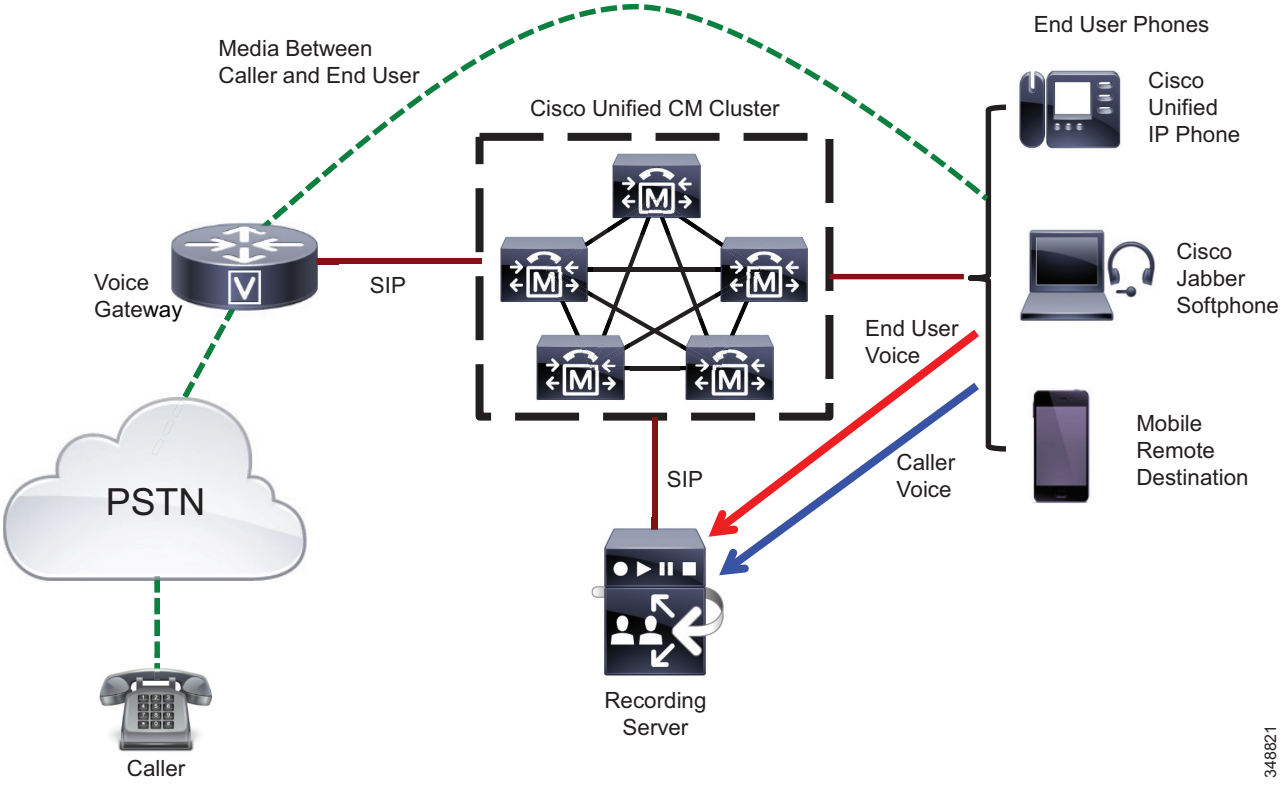
Cisco Unified CM network-based recording supports automatic and selective recordings for each individual line instance. This is accomplished by assigning a Recording Profile to each instance of a line where recording is required. This allows for recording on a single line of a multi-line device or a single instance of a shared line. In automatic recording, Unified CM automatically records every call that is connected on the endpoint. In selective recording, the user or an external application via JTAPI/CTI has to explicitly request Unified CM to start the recording for the call on the endpoint. Users can make the recording request by pressing the Start Recording button on the endpoint or by sending the recording request from the JTAPI or TAPI application. To start the recording, Unified CM sends the request to the forking device to fork the media of the conversation to the recording server, where the media is recorded.

Note If you have enabled both call recording and Multilevel Precedence and Preemption (MLPP), the lines that use both features will generate two additional call legs. Therefore, you must set the busy trigger for those lines to 3.

Unified CM Network-Based Recording with Built-in Bridge

Cisco Unified CM network-based recording with BIB uses the IP phone's built-in bridge to enable call recording. (See Figure 23-3.) During call recording, the agent phone forks the two streams to the recording server. The two streams, one for the called party's voice and one for the calling party's voice, get recorded separately. If a single stream is desired, customers can use third-party applications to mix the recorded streams to produce the conversation.

Figure 23-3 Unified CM Network-Based Recording Using a Phone's Built-in Bridge



For a list of Cisco Unified IP Phones that support call monitoring and recording with Unified CM, refer to the *Unified CM Silent Monitoring/Recording Supported Device Matrix*, available at

https://developer.cisco.com/site/uc-manager-sip/documents/supported/

Cisco Unified CM Network-Based Recording with a Gateway

When a call passes through a recording gateway, Cisco Unified CM network-based recording can utilize the gateway's media forking capability for call recording. When an external call is connected with an end user on the phone, Unified CM requests the gateway to fork the media of the conversations to the recording server through the UC Gateway Services API running on the gateway. The forked media consists of two RTP streams, one for end user voice and one for caller voice, and the recording server captures the streams separately. When a recording-enabled gateway is part of a call, several recording scenarios are possible, including external calls connected with end users on Cisco Unified IP Phones, Cisco Softphone (Cisco Jabber, for example) running on a PC, mobile phones as remote destinations, CTI ports, and Extend and Connect destinations. Essentially, once an external call terminates on the voice gateway that Unified CM is registered with, the entire conversation of the call from the caller's perspective can be recorded, no matter where the call goes inside the enterprise.

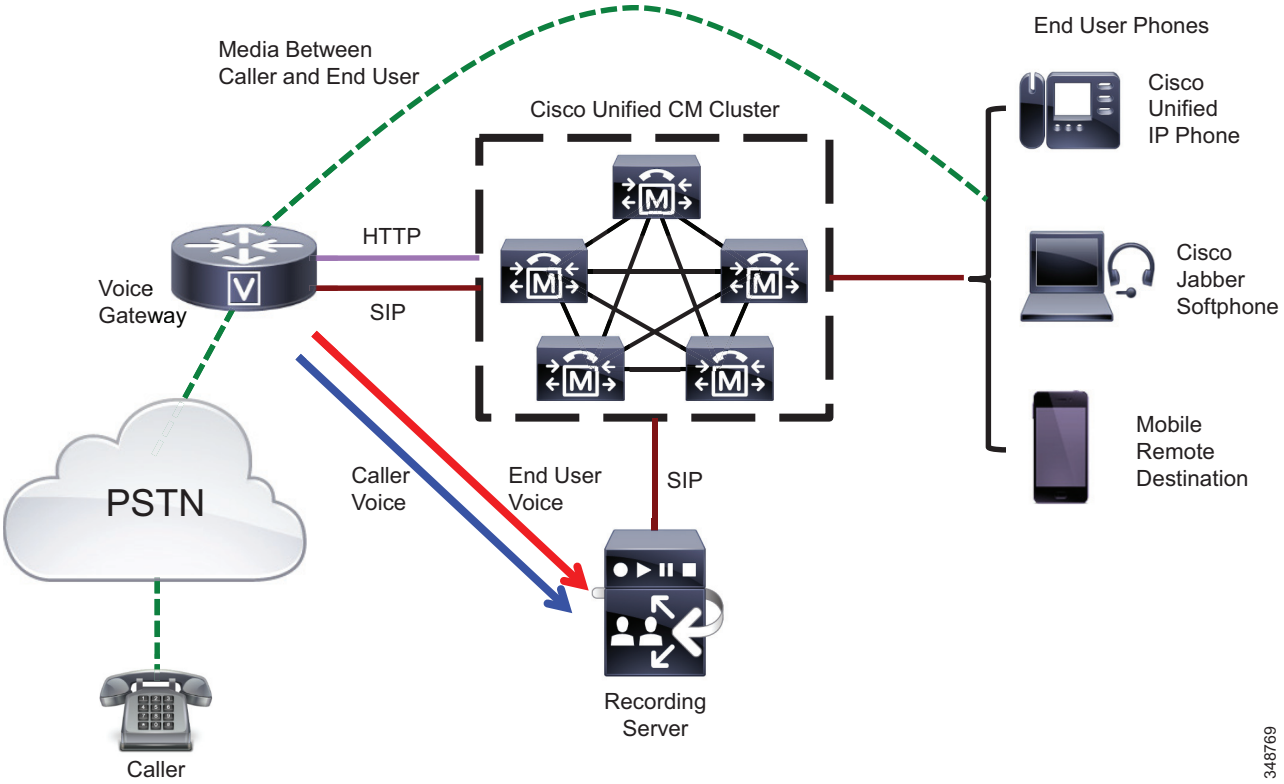
Cisco Unified CM network-based recording supports additional call types other than the ones described above. For details, refer to the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Note Invoking media forking from a voice gateway produces two RTP streams, and if silent monitoring is required, the application is responsible for mixing the streams.

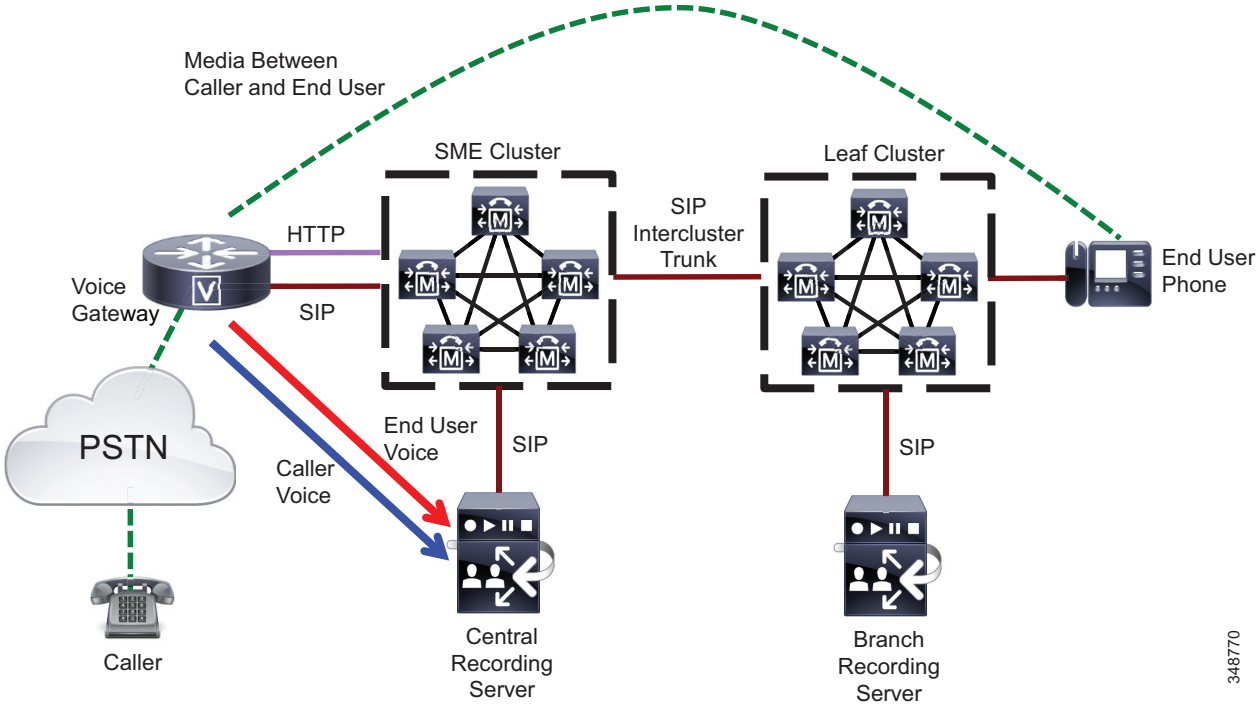
Figure 23-4 illustrates the basic setup for Cisco Unified CM network-based recording using gateways. Cisco Unified CM and the voice gateway are connected through a recording-enabled SIP trunk. Unified CM registers with the UC Gateway Services API running on the gateway through its HTTP interface. This enables Unified CM to receive call event notifications for all calls passing through the gateway and to decide when to start or stop the recording. Depending on the recording option configured, when a gateway call is connected with an end user on the phone, Unified CM might notify the gateway immediately to fork the media or wait for the user indication to start the recording before notifying the gateway. Unified CM notifies the gateway to stop forking the media upon user indication to stop the recording, or the gateway automatically stops the recording upon call termination. The requests to start or stop the recording are sent over the HTTP interface using the Extended Media Forking (XMF) API.

Figure 23-4 Network-Based Recording with a Gateway



With Unified CM network-based recording with a gateway, the end user phone and the media forking device (voice gateway) are decoupled. They can register to the same Unified CM cluster (as shown in Figure 23-4) or to separate Unified CM clusters. Therefore, this solution could be deployed in a multi-cluster environment such as Cisco Unified CM Session Management Edition (SME). Figure 23-5 illustrates an example of deploying Unified CM network-based recording with SME, where the voice gateway registers to the SME cluster and the end user phone registers to the leaf cluster. The SME cluster and leaf cluster are connected by a SIP intercluster trunk (ICT) with the gateway recording option enabled on both sides. Thus, the recording invocation requests and responses can be sent between SME and leaf clusters. Also, customers have the option to deploy the recording server centrally in the SME cluster with the voice gateway or to distribute the recording servers in all the leaf clusters.

Figure 23-5 Cisco Unified CM Network-Based Recording Deployment with SME



When deploying Unified CM network-based recording with a gateway, observe the following guidelines:

- Network-based recording with a gateway is supported on a variety of platforms including Cisco Integrated Services Routers (ISRs) (for example, ISR 4K) and Cisco Aggregation Services Routers (ASRs). For detailed requirements, refer to the latest version of the *Feature Configuration Guide for Cisco Unified Communications Manager*, available at

  https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

- Only SIP is supported between the voice gateway and Cisco Unified CM, and SIP proxy servers are not supported.
- For inter-cluster recording, only a SIP trunk is supported to interconnect the clusters.
- Secure recording is not supported.
- IPv6 is not supported.

Agent Desktop

Agent desktop monitoring and recording solutions are specific to contact center deployments that enable supervisors to do silent monitoring and initiate call recording when needed. Several agent desktop monitoring and recording solutions are available, such as:

- Cisco Agent Desktop (CAD) Silent Monitoring and Recording
- Cisco Remote Silent Monitoring (RSM)

These solutions are described in detail in the latest version of the following documents:

- *Solution Design Guide for Cisco Unified Contact Center Enterprise*, available at

  https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-implementation-design-guides-list.html
- *Solution Design Guide for Cisco Unified Contact Center Express*, available at

  https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-express/products-implementation-design-guides-list.html

Capacity Planning for Monitoring and Recording

Enabling any type of monitoring and/or call recording impacts the overall Unified Communications system capacity. Some silent monitoring and recording solutions (such as the silent monitoring and recording feature based on Unified CM) consume resources from Unified CM, whereas other solutions such as SPAN or desktop silent monitoring and recording do not. Consider the following points when doing capacity planning for Unified Communications systems with call recording enabled:

- With Unified CM call recording, each recorded call adds two calls to the call processing component BHCA capacity. Forking media from an IP phone or voice gateway consumes resources from Unified CM or the voice gateway, respectively.
- Bandwidth requirements increase when media forking is enabled on IP phones or Cisco Unified Border Element devices to send forked media to the recording server. In case of agent desktop monitoring and recording, the bandwidth utilization can be bursty, depending on how many calls are being monitored or recorded at a given time.
- Call recording using a Cisco Unified Border Element doubles the weight of a call. Thus, call capacity would be cut in half if all calls passing through the Cisco Unified Border Element were recorded.
- Memory utilization on Cisco Unified Border Element increases for each call that is recorded.
- In cases where CTI applications interact with Cisco Unified CM to invoke recording and monitoring, you should consider the Unified CM cluster deployment model and load-balance the CTI applications across the cluster.

Due to the complexity associated with sizing, all deployments must be sized with the Cisco Collaboration Sizing Tool, available to Cisco employees and partners only (with proper login authentication) at

https://cucst.cloudapps.cisco.com/landing

PART 4

Collaboration System Provisioning and Management

Contents of This Part

This part of the document contains the following chapters:

- Overview of Collaboration System Provisioning and Management
- Collaboration Solution Sizing Guidance
- Cisco Collaboration System Migration
- Network Management

Overview of Collaboration System Provisioning and Management

Revised: June 15, 2015

Once the network, call routing, call control infrastructure, and applications and services have been put in place for your Cisco Unified Communications and Collaboration System, network and application management components can be added or layered on top of that infrastructure. There are numerous applications and services that can be deployed in an existing Cisco Unified Communications and Collaboration infrastructure to monitor and manage the operations of the system. These applications and services can be classified into four basic areas:

- User and device provisioning services — Provide the centralized ability to provision and configure users and devices for Unified Communications and Collaboration applications and services.
- Voice quality monitoring and alerting — Provide the ability to monitor on an ongoing basis various call flows occurring within the system to determine whether voice and video quality are acceptable and to alert administrators when the quality is not acceptable.
- Operations and fault monitoring — Provides the centralized ability to monitor all application and service operations and to issue alerts to administrators regarding network and application failures.
- Network and application probing — Provides the ability to probe and collect network and application traffic information at various locations throughout the deployment and to allow administrators to access and retrieve this information from a central location.

This part of the SRND covers the applications and services mentioned above. It provides an introduction to the various network management applications and services, followed by discussions surrounding architecture, high availability, capacity planning, and design considerations. The discussions focus on design-related aspects of the applications and services rather than product-specific support and configuration information, which is covered in related product documentation.

This part of the SRND also contains detailed information on how to size a Cisco Unified Communications and Collaboration deployment as well as some recommended methods for migrating from third-party and legacy communications systems to a Cisco Unified Communications and Collaboration System.

This part of the SRND includes the following chapters:

- Collaboration Solution Sizing Guidance, page 25-1

  This chapter discusses the sizing of individual Unified Communications and Collaboration components as well as systems consisting of several components communicating with each other. This chapter also discusses the performance impact of the different functions that the various Unified Communications and Collaboration products support, and it explains why "designing by datasheets" is not the preferred way to deploy a complex Unified Communications and Collaboration network. In addition, this chapter provides insights on how to work with the various sizing tools available, mostly notably the Cisco Collaboration Sizing Tool.

- Cisco Collaboration System Migration, page 26-1

  This chapter describes several methods for migrating from separate standalone voice, video, and collaboration systems to an integrated Cisco Unified Communications and Collaboration System. It discusses the pros and cons of both phased migration and parallel cutover. It also describes the services needed to connect a private branch exchange (PBX) to a new Unified Communications and Collaboration system. The major topics discussed in this chapter include IP telephony migration, video migration, and migration of voice and desktop collaboration systems.

- Network Management, page 27-1

  This chapter examines Unified Communications and Collaboration network and application management services, a common and prevalent set of services within most Unified Communications and Collaboration deployments, which allow administrators to provision and configure users and devices, monitor network and application operations as well as voice and video quality, and receive alerts and alarms when issues arise. This chapter also examines the impact of these management applications and services on deployment models and provides design and deployment best practices for network and application management services and applications.

Architecture

As with other network and application technology systems, operations and serviceability applications and services must be layered on top of the underlying network, system, and application infrastructures in order to be able to monitor and control those infrastructures. Unified Communications and Collaboration operations and serviceability services such as user and device provisioning, voice and video quality monitoring and altering, operations and fault monitoring, and network and application probing, all rely on the underlying network infrastructure for network connectivity for various operations and serviceability applications and probes. While there is no direct reliance on the Unified Communications and Collaboration call routing, call control infrastructure, or Unified Communications and Collaboration clients and services, these infrastructures and applications are what the various operational and management services actually manage and configure. For example, user and device provisioning services as well as various monitoring and alerting services leverage the network infrastructure for connectivity to various Unified Communications and Collaboration applications and service nodes in order to configure and monitor various components and operations. These same services also communicate directly with, and in some cases change configurations on or receive alerts from, components such as call processing agents, PSTN and IP gateways, media resources, endpoints, and various Unified Communications and Collaboration applications for messaging, rich media conferencing, and collaboration clients. In addition to relying on these infrastructure layers and basic Unified Communications and Collaboration services and applications, services pertaining to operations and serviceability are also often dependent upon each other for full functionality.

High Availability

As with network, call routing, and call control infrastructures and critical Unified Communications and Collaboration applications and services, operations and serviceability services should be made highly available to ensure that required provisioning, monitoring, and altering will continue even if failures occur in the network or applications. It is important to understand the various types of failures that can occur as well as the design considerations around those failures. In some cases, the failure of a single operations and management application or server can impact multiple services because the Unified Communications and Collaboration operations and serviceability components are dependent on other components or services. For example, while the various application service components of a network management deployment might be functioning properly, the loss of network connectivity to, or a failure of, a network probe would effectively eliminate the ability to monitor network health or voice and video quality unless redundant network probes have been deployed along with alternate paths of connectivity.

For operations and serviceability functions such as user and device provisioning, high availability considerations include temporary loss of functionality due to network connectivity or application server failures resulting in the inability of administrators to provision users and devices or to make changes to those user accounts or device configurations. In addition, failover considerations for these types of operations include scenarios in which portions of the functionality can be handled by a redundant operation or management application that allows administrators to continue to facilitate some configuration changes in the event of certain failures.

High availability considerations are also a concern for operations and serviceability applications that provide services such as voice and video quality monitoring or application and operations fault monitoring. Interrupted network connectivity or server or application failures will typically result in a reduced ability to monitor and/or alert, and in some cases will cause complete loss of such functionality. For voice and video quality monitoring, this can mean that quality measurements for some call flows or devices will be unavailable. For operations and fault monitoring services, high availability considerations include the potential for loss of operational change tracking data or fault alerts and indications.

Capacity Planning

Network, call routing, and call control infrastructures as well as Unified Communications and Collaboration applications and services must be designed and deployed with an understanding of the capacity and scalability of the individual components and the overall system. Similarly, deployment of operations and serviceability components and services must also be designed with attention to capacity and scalability considerations. When deploying various operations and serviceability applications and components, not only is it important to consider the scalability of these applications themselves, but you must also consider the scalability of the underlying infrastructures. Certainly the network infrastructure must have available bandwidth and be capable of handling the additional traffic load that those operations will create. Likewise, the call routing and control infrastructure must be capable of handling required inputs and outputs as facilitated by the various operations and serviceability components in use. For example, with operational applications and services such as voice quality monitoring and alerting and operations and fault monitoring, there are capacity implications for each of these individual applications or services in terms of the number of devices and call flows that can be monitored at a given time, but just as important is the scalability of the underlying infrastructure and monitored applications to handle the added network traffic and connections required for monitoring and alerting. While the monitoring and alerting application or service itself may be able to support the monitoring of many network devices and call flows, the underlying network or devices might not have available capacity to handle the probing connections or the alarm messaging load generated by the monitoring and alerting services.

For operation applications or services that provide user or device provisioning capabilities, capacity planning considerations include things such as ensuring that the provisioning application can handle the requested load and also that user or device provisioning operations not only do not exceed the number of support devices or users for a particular underlying Unified Communications application or service, but also that provisioning or configuration change transactions do not exceed either the capacity of the underlying network or the rate at which a particular application can handle transactions. In most cases additional capacity can be added by increasing the number of operational provisioning application servers or by increasing the size or number of underlying Unified Communications and Collaboration applications or service instances, assuming the underlying network and call routing and control infrastructures are capable of handling this additional load.

For a complete discussion of system sizing, capacity planning, and deployment considerations related to sizing, refer to the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Collaboration Solution Sizing Guidance

Revised: March 1, 2018

This chapter describes system sizing for Cisco Collaboration products and systems. Sizing involves providing an accurate estimate of the required hardware platforms for the system, based on the number of users, traffic mix, traffic load, and features that the system will provide.

Accurate sizing is critical to ensure that the deployed system will meet the expected service quality for call volumes and throughput. For standalone products, manual calculation of the system size may be feasible (as covered in the section on Sizing for Standalone Products, page 25-49). However, there are many sizing factors to consider in a complex system deployment. For example, multiple products may be distributed across different locations and may include video endpoints, call centers, and voice/video conferencing. Cisco Systems provides a set of sizing rules to handle the resulting complexity.

This chapter provides a general introduction to system sizing methodology and the factors that affect sizing, and also provides information about how to use the sizing tools.

Note This chapter should be read in conjunction with the product descriptions and design and deployment considerations covered in other chapters of this document. A good understanding of both of these aspects is required for a successful deployment.

This chapter includes the following major sections:

- What's New in This Chapter, page 25-2
- Methodology for System Sizing, page 25-2
- System Sizing Considerations, page 25-9
- Sizing Tools Overview, page 25-10
- Using the SME Sizing Tool, page 25-12
- Using the VXI Sizing Tool, page 25-13
- Using the Cisco Collaboration Sizing Tool, page 25-13
- Sizing for Standalone Products, page 25-49

Note For simplified sizing guidance without the use of the Collaboration Sizing Tool, refer to the latest version of the *Cisco Preferred Architecture for Enterprise Collaboration CVD*, available at https://www.cisco.com/go/pa.

What's New in This Chapter

Table 25-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 25-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Sizing for Cisco Jabber clients	Cisco Jabber Clients, page 25-18	March 1, 2018
Sizing for centralized IM and Presence clusters	Centralized IM and Presence, page 25-35	March 1, 2018

Methodology for System Sizing

To ensure accurate system sizing, Cisco follows a methodology that is supported by actual performance test results and that incorporates industry-standard traffic engineering models to estimate the maximum expected traffic that the system needs to handle during normal operating conditions.

The following sections describe the sizing methodology:

- Performance Testing, page 25-2
- System Modeling, page 25-3
- Traffic Engineering, page 25-5

Performance Testing

Each product performs a set of functions, and each function utilizes a number of resources (such as CPU and memory). Cisco defines and executes performance tests that allow us to measure resource usage accurately for each function at different usage levels.

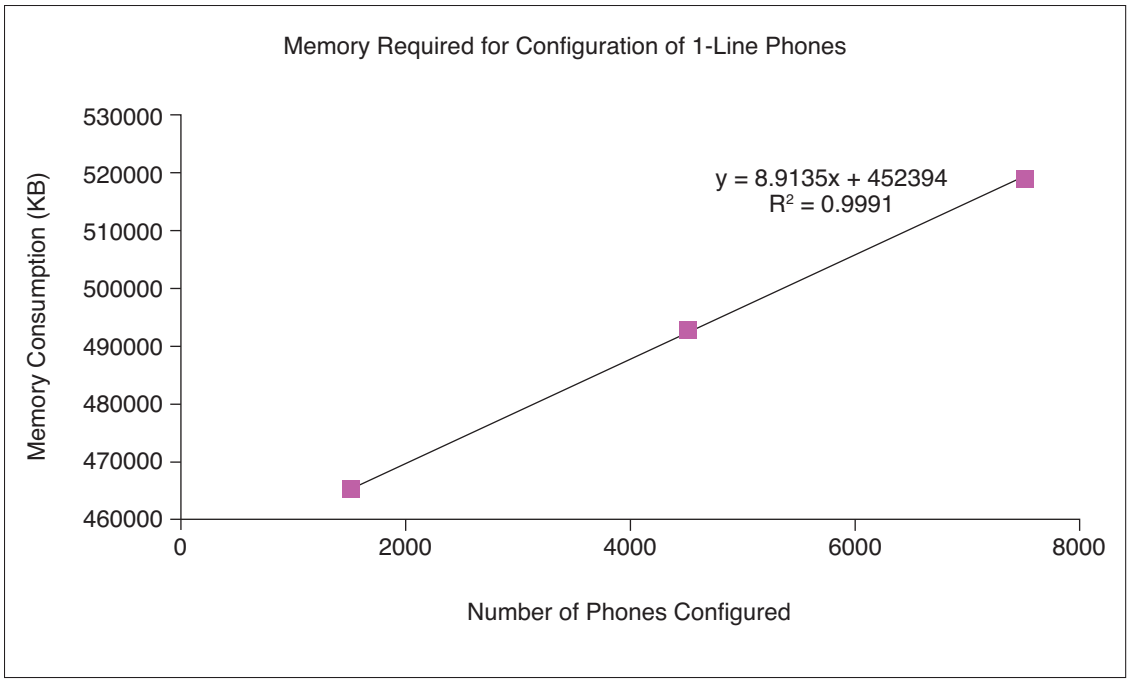
Most systems exhibit linearity within a certain range, beyond which the system performance can become unpredictable. Cisco sets the usage levels for each performance test to identify and confirm the linear range of the resource usage for each function. The results for each test can be graphed using a minimal number of data points. If required, additional data points (at intermediate load levels) are obtained in order to define the actual system behavior.

The slope of the linear section of the graph defines the resource usage and/or cost for each incremental addition of work. The R^2 value is used to estimate the closeness of the fit. If the R^2 value is close to 1, the formula is a close match for the data.

For example, Figure 25-1 shows the results of a test conducted to determine the memory requirements for configuring single-line IP phones. It shows the memory consumed by configuring 1500, 4500, and 7500 single-line IP phones in Unified CM. The graph shows that the equation of the trend line is linear and can be used to predict the dependent variable (in this case, memory) based on the control variable (the number of phones).

In this particular test, the R^2 value is extremely close to 1. From the equation, we can compute that the memory consumed with configuration of 7,500 one-line phones is approximately 519,000 Kbytes and that each additional line configured for an endpoint in the system consumes an additional 8.91 Kbytes.

Figure 25-1 Memory Required for Configuration of One-Line Phones



System Modeling

Cisco uses the performance test results to create a system model. A system model is a mathematical model that calculates the maximum resource usage for a specified set of features, endpoints, and traffic mix, which are provided as inputs to the model.

To develop a system model for a given product, Cisco performs the following steps:

1. Itemize all of the functions that the product performs. Identify variations of the function that need to be tested. For example, each type of call will potentially use a different amount of the measured resources.
2. Determine the resources of interest. Generally this includes memory and CPU. Specific products may have additional resources that impact system sizing.
3. Run the performance tests (as described in the previous section) to determine the resource usage for each function.
4. For each function, use the linear range to define the formula for resource usage.

We may need to repeat these steps a number of times because other factors (such as software release, call mix, and types of endpoints) can impact resource usage.

The system model for the product consists of aggregating the formulas for each function supported by the product. The model can be fairly simple for some products, but it can be very complex for a product that supports multiple functions, multiple endpoint types, and multiple call types.

Specific considerations for memory and CPU resource types are described in the following sections.

Memory Usage Analysis

The system model differentiates between static and dynamic memory, which have different usage characteristics. There is also system memory, which is reserved for the operating system and other processes. These three memory types are described in the following sections:

Static memory

Static memory is consumed even when there is no traffic on the system. Static memory usage includes the data for system configuration and the data for registered endpoints. Static memory also includes configuration for the dial plan (which covers items such as partitions, translation patterns, route lists and groups). In addition, static memory includes the memory allocated for CTI and other applications. In a large system, static memory is mainly a function of the number of configured endpoints and the size of the dial plan.

Note that each type of endpoint may consume a different amount of memory. Memory usage may also depend on the device protocol (SIP or SCCP), the number of line appearances, security capabilities, and other factors. Each of these variants must be measured and incorporated into the model.

Dynamic memory

Dynamic memory is used for transient activities, such as saving the context of each active call. In a large system, dynamic memory is primarily a function of the number of concurrent calls.

The number of concurrent calls is determined by the average call holding time (ACHT). A longer ACHT results in more dynamic memory use because there will be a larger number of concurrent active calls.

Memory usage may vary considerably for different types of calls and different protocols (such as SCCP and SIP).

System memory

System memory is required by the operating system (OS) and by other processes and services. In addition, some memory may be reserved for transient spikes in usage. System memory reduces the amount of memory available for applications running on the platform.

CPU Usage Analysis

An inactive system exhibits some CPU activity, but most of the CPU utilization occurs during transaction processing, such as setting up and tearing down calls. Therefore, one of the key determinants of CPU usage is the offered call rate.

CPU usage can vary considerably depending on the type of calls. Calls can originate and terminate within the same server, or they can originate and terminate on two different servers or clusters. Calls can also originate from the Unified CM cluster and terminate to a PSTN gateway or trunk.

CPU usage analysis must account for the different cost of a call originating versus terminating on Unified CM, the protocols in use, and whether security features are enabled. CPU usage also depends on factors such as the configuration database complexity and whether CDRs or CMRs are being generated.

CPU usage will vary substantially depending on the actual hardware platform. Therefore, the same performance tests must be repeated on all platforms that are supported for each product.

CPU usage is also affected by CPU-intensive call operations such as call transfers, conferences, and media resource functions such as MTP or music on hold. Shared lines consume additional CPU resources, because each call to a shared line is offered to all of the phones that share the line.

Traffic Engineering

Cisco uses industry-standard traffic engineering models to estimate the dynamic load on the system.

Traffic engineering provides mathematical models that calculate the maximum traffic level expected for a set of users. The models also determine the amount of a shared resource (such as PSTN trunks) that is required to support a given traffic load.

The following sections describe traffic engineering considerations for different types of traffic:

- Definitions, page 25-5
- Voice Traffic, page 25-6
- Contact Center Traffic, page 25-7
- Video Traffic, page 25-7
- Conferencing and Collaboration Traffic, page 25-8

Definitions

Traffic engineering defines the following terms:

Maximum Simultaneous Calls

The maximum number of simultaneous active calls that the system can handle at one time.

Calls per Second

The number of new call attempts that arrive at the system in one second, plus the number of existing calls torn down during that same one second interval. This unit can be used to define the average calls per second that the system expects to handle during a busy hour. (This number is equivalent to the busy hour calls divided by 60.)

This unit can also be used to define the maximum burst of traffic that the system needs to handle.

Busy Hour

The hour in a given 24-hour period during which the maximum total traffic occurs. This hour varies depending on the organization and the type of traffic. For business voice traffic, the busy hour is traditionally assumed to be during morning hours (for example, 10 AM to 11 AM).

Busy Hour Call Attempts (BHCA)

The user BHCA represents the average number of calls that a user initiates or receives during the busy hour. Typically, BHCA will be calculated as the average of the busy hour call attempts from the busiest 30 days of the year). System BHCA is the User BHCA multiplied by the number of users.

Blocking Factor

Indicates a grade of service, expressed as the probability that a call will be blocked during the busy hour due to lack of resources. For example, a blocking factor of 1% indicates that one out of every 100 calls may be blocked due to lack of resources required to process the call.

Average Call Hold Time

This is the average period of time that the resource is busy. For example, on a voice call the ACHT is the period of time between call setup and call tear-down when there is an open speech path between the two parties. A hold time of 3 minutes (180 seconds) is an industry average used for traffic engineering of voice systems.

Erlang

The Erlang is a measure of traffic load on a system. To calculate Erlangs, multiply calls per hour by the average holding time (in hours). Resource requirements can be derived from Erlangs by using the appropriate Erlang model.

The number of Erlangs handled by a resource (such as a trunk group) is equal to the number of simultaneous calls. The Erlang value is usually averaged over a one-hour period of time.

Erlang B Model

The Erlang B model can determine the number of trunks required to handle a traffic load (in Erlangs) with a specified blocking factor. The Extended Erlang B model includes the modeling of retries (for calls that are blocked). The retry percentage is an additional input to the Extended Erlang B model.

Erlang C Model

The Erlang C model incorporates queuing of incoming calls, and is therefore very useful for modeling call center traffic.

Bursty Traffic

Traffic models assume a fairly steady arrival rate for the call attempts, which is a valid assumption for a large number of subscribers acting independently. However, in a real system, a number of calls could arrive over a very short period of time. Such a traffic burst will consume the system resources very quickly, and can result in a high number of blocked calls. Products may specify the size and duration of traffic bursts that they can handle.

Voice Traffic

Standard voice traffic is characterized by specifying the busy hour call attempts (BHCA) and the average call holding time (ACHT). For example, if the system BHCA is 200 and the average call duration is 3 minutes, the system is being used for a total of 600 minutes, which is 10 Erlangs.

To calculate the usage of a shared resource (such as a PSTN trunk group), the blocking factor must also be specified. For example, given an Erlang value and the blocking factor, we can use an Erlang calculator or lookup tables to calculate the number of voice circuits that will be required on PSTN gateways.

Table 25-2 illustrates the relationship between number of trunks, blocking probability, and Erlangs of traffic.

Table 25-2 Erlang B Traffic Table (Number of Circuits Required)

Number of Erlangs	Blocking Probability					
	0.05%	1%	2%	3%	4%	5%
10	19	18	17	16	15	15
20	32	30	28	27	26	26
30	44	42	39	38	37	36

From Table 25-2 we can determine the following information:

- Given an Erlang requirement of 20 and a blocking factor of 1%, the system will need 30 circuits.
- Additional circuits are required to provide a lower blocking factor (such as 1%) than to provide a higher blocking factor (such as 5%).

Contact Center Traffic

Contact centers demonstrate a unique pattern of traffic, because these systems typically handle large volumes of calls that are handled by a small number of agents or interactive voice response (IVR) systems. Contact centers are engineered for high resource utilization, therefore their agents, trunks, and IVR systems are kept busy while they are in operation, which usually is 24 hours a day. Call queuing is typical (when incoming call traffic exceeds agent capacity, calls wait in queue for the next available agent), and the agents are usually dedicated during their work shifts to taking contact center calls.

Average call holding times for contact centers are often shorter than for normal business calls. Many calls interact only with the IVR system and never need to speak to a human agent. These calls are known as self-service calls. The average holding time for self-service calls is about 30 seconds, while a call serviced by an agent may have an average holding time of 3 minutes (the same as normal business traffic), making the overall average holding time in the contact center shorter than for normal business traffic.

The goal of contact center management is to optimize resource use (including IVR ports, PSTN trunks, and human agents), therefore resource utilization will be high.

A call center usually has a higher call arrival rate than a typical business environment. These call arrival rates can also peak at different times of day (not the usual busy hours) and for different reasons than normal business traffic. For example, when a television advertisement runs for a particular holiday package with a 1-800 number, the call arrival rate for the system will experience a peak of traffic for about 15 minutes after the ad airs. This call arrival rate can exceed the average call arrival rate of the contact center by an order of magnitude.

As noted earlier, contact center sizing uses the Erlang C model to account for calls waiting in queues. Contact centers require additional resources, such as interactive voice response (IVR) ports. The time that calls wait in queues needs to be factored in when sizing the PSTN gateways (see Gateway Sizing for Contact Center Traffic, page 25-39).

Note For additional information about Cisco Unified Contact Center deployments, refer to the latest version of the *Solution Design Guide for Cisco Unified Contact Center Enterprise*, available at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-implementation-design-guides-list.html.

Video Traffic

Point-to-point video traffic demonstrates similar characteristics to its voice equivalents for call arrival rates, peak usage times, and call durations. Also, signaling for call setup and take-down is similar to voice calls.

Video traffic requires significantly higher network bandwidth than voice because the payload in video packets is much larger than in voice packets. Also, video traffic can be much burstier than voice. Voice packet sizes are usually fairly consistent (specifics depend on the encoding algorithm in use), whereas video frames can vary considerably in size, depending on how much change has occurred since the previous frame. The resulting RTP packet stream can therefore exhibit bursts of traffic.

Implications for video conferencing are covered in the next section.

Conferencing and Collaboration Traffic

Conferencing traffic has considerably different characteristics than point-to-point voice/video calls. The traffic model for conferencing traffic needs to accommodate the following differences:

- Call arrivals

  A traditional traffic model assumes a Poisson distribution of busy-hour call arrivals throughout the busy hour. However, most participants join their conference call within 5 to 10 minutes of the meeting start time, and most conference calls are scheduled to start at the beginning of the hour. Therefore, the call arrival rate will exhibit a single burst at the top of the hour rather than a Poisson distribution throughout the hour.

- Peaks

  Business voice traffic typically has a distinct peak in the morning (between 10:00 and 11:00 AM) and another peak in the afternoon (between 1:00 and 2:00 PM). However, conference facilities are generally a limited resource, resulting in meetings that are distributed more evenly throughout the business day, with less of a pronounced peak at peak times.

- Call durations

  The average business voice call duration is 3 minutes. The average conference call duration may be closer to 50 minutes (depending on the mix of 30 minute, 60 minute, and longer meetings).

- Video conferencing

  Specialized equipment is required to provide the switching or combining of video streams. Therefore, expected usage of video endpoints is an important factor in the model.

Sizing a deployment for conferencing primarily involves deciding how many concurrent connections are required. For example, sizing for TelePresence Servers would include the following considerations:

- Geographical location — Each region served by Unified CM should have dedicated conferencing resources.
- Preference for TelePresence Server platforms — Hardware or software
- TelePresence Server platform capacities
- TelePresence Conductor platform capacities
- Type of conferencing — Audio and/or video; scheduled and/or non-scheduled
- Conference video resolution — Higher quality conferences use more resources.
- Large conference requirements — For example, all-hands meetings

Conference resources are generally dedicated to a region in order to keep as much of the conference media on the regional network; therefore, sizing can be considered on a region-by-region basis.

System Sizing Considerations

For large and complex deployments, the system designer will need to consider a number of design and deployment factors that influence system sizing. These factors are described in the following sections:

- Network Design Factors, page 25-9
- Other Sizing Factors, page 25-10

Network Design Factors

Solution sizing is affected by the following network design factors:

- Cluster sizes

  A major design decision is whether to create a large centralized Cisco Unified CM cluster or to create a cluster at each major location. The central cluster may have a higher utilization, but you may be forced into a second cluster if a cluster limit is exceeded.

  Some system limits are not absolute and can change dynamically based on the sizing of other services configured in the system.

- Interaction between individual products

  Unified CM plays a central role in most Cisco Collaboration deployments, and it is affected by other components in the system. For example, the addition of Cisco WebEx Meetings Server will tend to concentrate a large number of call setups into a short period (at the beginning of conferencing sessions). Depending on the other functions covered by Unified CM, this may require the addition of Unified CM server nodes.

- Server capabilities

  Each type of server or router supports different capabilities. For example, more powerful servers might have a higher number of network ports compared to Cisco Business Edition 6000 platforms or a Cisco Integrated Services Router (ISR).

  As another example, different models of Cisco Integrated Services Routers have restrictions on the number and types of network modules or Cisco Unified Computing System (UCS) E-Series blade servers they can host.

- Optional capabilities and features

  The system sizing can be impacted if you enable options such as call detail recording (CDR) or call management record (CMR) generation.

Other Sizing Factors

The following additional factors also affect system sizing:

- Mix of call types:

  There are variations in resources consumed by each call type: calls between phones in the same subscriber node, calls between two subscriber nodes in the same cluster, calls between two clusters, and calls that flow to and from the PSTN. Even calls from different types of phones and gateways are different, depending on the protocol and services such as video.

- Mix of endpoint types

  The expected number of phones and users is another example of an obvious factor that affects sizing. Here again, the type of phones, the number of lines configured on them, and whether they are in secure mode, among other things, have an impact on system sizing.

- System release

  System resource usage can vary between system releases. Sometimes, new capabilities in a release can cause an increase in resource usage. In other cases, software improvements can result in a decrease in resource usage.

- Use of external applications

  External applications can communicate with the call processing agent by using an interface such as CTI. This load needs to be factored into the system sizing.

- Anticipated system growth

  If system usage is expected to grow in the next year or two, it would be preferable to build that growth into the original system rather than face a potentially disruptive upgrade in the near future.

- Average and peak usage

  Ensure that the system sizing is based on a realistic view of peak usage. If the peak is underestimated, the system could experience service degradation or equipment outages when the actual peak traffic is encountered.

Because of all the factors and possible variations, the accurate sizing of a large system deployment is a complex undertaking. For this reason, Cisco strongly recommends using the system sizing tools described in the following sections.

Sizing Tools Overview

Cisco provides several sizing tools to assist with accurate solution sizing. The sizing tools are available at the following location (only Cisco employees and certified partners can access this site):

https://cucst.cloudapps.cisco.com/landing

Cisco recommends that you use the sizing tools to perform system sizing. These tools take into account data from performance testing, individual product limits and performance ratings, advanced and new features in product releases, design recommendations from this SRND, and other factors. Based on input provided by the system designer, the tools apply their sizing algorithms to the supplied data to recommend a set of hardware resources.

If you do not have access to the sizing tools, please contact your Cisco account representative or Cisco partner integrator to obtain system sizing information.

Tool-specific sections below contain explanations of the inputs required for the tool and how the inputs can best be collected from an existing system or estimated for a system still in the design stage. Obviously, the sizing recommendations generated by the tools are only as accurate as the input data you provide.

Cisco provides the following sizing tools:

- Cisco Collaboration Sizing Tool

  This tool guides the user through a complete system deployment. The tool covers the following products and components:

  - Cisco Unified Communications Manager (Unified CM)
  - IM and Presence services
  - Voice messaging
  - Conferencing
  - Gateways
  - Cisco Unified Communications Management Suite
  - Cisco Unified Contact Center components

- Cisco Unified Communications Manager Session Management Edition (SME) Sizing Tool

  This is a specialized tool that focuses on the specific functions of a Unified CM Session Management Edition deployment.

- Cisco VXI Sizing and Configuration Tool

  This is a specialized tool for sizing the Cisco Virtual Experience Infrastructure (VXI).

For more information on these tools and their access privileges, refer to the *Collaboration Sizing Tool Frequently Asked Questions*, available at

https://cucst.cloudapps.cisco.com/help/UC_Sizing_Tools_FAQ.pdf

Caution If any parameter of your system design exceeds the range of values that the above sizing tools allow you to enter, consult your Cisco account team or a Cisco Systems Engineer (SE) about your design before proceeding further.

In addition to these sizing tools, a Virtual Machine Placement Tool is available to Cisco partners and customers with a valid login account. The Virtual Machine Placement Tool is a graphical tool that allows you to select Tested Reference Configurations (TRC) or specifications-based hardware, and to drag and drop the various Cisco Collaboration application virtual machines on those servers. Some placeholders representing third-party application virtual machines are also available when deploying Cisco Collaboration applications co-resident with third-party applications. The sizing tools determine how large the servers need to be and how many virtual machines are necessary. This information can then be entered as an input to this Virtual Machine Placement Tool in order to determine how to place the various virtual machines and to determine how many servers would need to be deployed. Even though some of the co-residency rules are implemented in the tool, Cisco recommends verifying the rules by using the guidelines documented at

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/collaboration-virtualization-sizing.html

The Virtual Machine Placement Tool is available (with proper login authentication) at

https://www.cisco.com/go/vmpt

Using the SME Sizing Tool

The Session Management Edition (SME) is a Unified CM operating in a specific deployment mode. In a pure SME deployment, call traffic runs only across trunk interfaces and the SME hosts no line interfaces.

An SME cluster follows the same topology as a regular Unified CM cluster. A publisher node provides the master configuration repository. The TFTP service can run on the publisher node if the number of phones or MGCP gateways in the cluster is relatively small. A redundancy ratio of 1:1 is recommended for call processing subscribers.

To size an SME cluster, you must consider the functionality that it is expected to perform. In a base configuration, the SME acts as a routing aggregation point for a number of leaf clusters. It also provides centralized PSTN access for all of the leaf clusters connected to it. In more advanced configurations, the SME may also host centralized voice messaging, mobility, and conferencing services. The performance of the SME is influenced by the type of trunk protocols that the leaf clusters use to connect to it and by the BHCA across those trunks.

The SME sizing tool requires the following input parameters:

- The various types of trunk interfaces that the cluster services. The following trunk protocols are supported by the SME; however, Cisco recommends SIP trunks as the preferred protocol:
  - SIP
  - H.323
  - MGCP (Q.931)
  - SIP (Q.SIG)
  - H.323 Annex M1
  - MGCP (Q.SIG)
- The number of users that access SME cluster services through each type of trunk interface
- BHCA per user for each trunk interface to leaf clusters for intercluster calls
- BHCA per user for each trunk interface to leaf clusters for off-net (PSTN) calls
- The type of trunk interface used by the SME cluster to connect to the PSTN
- Average holding time for calls
- Number of route and translation patterns

If the SME acts as a service aggregation point, you must consider the following additional sizing parameters:

- For centralized voice messaging, the percentage of calls that are sent to voice mail
- For mobility, the number of users and the remote destinations per user
- For conferencing service, the conferencing dial-in interval

The performance of the SME is measured as calls-per-second across each pair of protocols. There are variations across the hardware platforms and software versions.

Using the VXI Sizing Tool

Cisco Virtualization Experience Infrastructure (VXI) is a systems approach that unifies virtual desktops, voice, and video, to provide a superior virtual workspace experience. The Cisco VXI Sizing Tool assists with the task of sizing components for a Virtualization Experience Infrastructure solution.

Using the Cisco Collaboration Sizing Tool

The Cisco Collaboration Sizing Tool covers sizing for a number of products and components. For a complete list of components and versions supported by tool, see the release notes that are included in the sizing tool installation package.

The following sections describe the significant factors that influence sizing of the individual products and also how these individual products can influence the sizing considerations of other products in the system deployment:

- Cisco Unified Communications Manager, page 25-13
- Media Resources, page 25-28
- Cisco Unified CM Megacluster Deployment, page 25-32
- Cisco IM and Presence, page 25-33
- Emergency Services, page 25-36
- Gateways, page 25-38
- Voice Messaging, page 25-42
- Collaborative Conferencing, page 25-44
- Cisco Prime Collaboration Management Tools, page 25-48
- Cisco Unified Communications Manager Express, page 25-49
- Cisco Business Edition, page 25-49

Cisco Unified Communications Manager

Cisco Unified Communications Manager (Unified CM) is the hub of any Unified Communications deployment. It performs key functions such as controlling endpoints, routing calls, enforcing policies, and hosting applications. Unified CM provides coordination for the other Unified Communications products such as PSTN gateways, Cisco Unity Connection, Cisco Unified Communications Manager IM and Presence Service, and Cisco Unified Contact Center. The coordination function has an impact on Unified CM performance, and therefore must be accounted for in Unified CM sizing.

A number of factors affect Unified CM performance and must be considered when sizing a Unified CM deployment. These factors are described in the following sections:

- Virtual Nodes and Cluster Maximums, page 25-14
- Deployment Options, page 25-14
- Endpoints, page 25-16
- Cisco Collaboration Clients and Applications, page 25-17
- Call Traffic, page 25-22
- Dial Plan, page 25-23

- Applications and CTI, page 25-23
- Media Resources, page 25-28

Virtual Nodes and Cluster Maximums

The sizing tool applies the following server node and cluster maximums. These values can vary depending on Unified CM software version:

- Each cluster can support configuration and registration for a maximum of 40,000 secured or unsecured SCCP or SIP phones.
- Two TFTP server nodes are required, in addition to a dedicated publisher, if the number of endpoints in the cluster exceeds 1,250.
- Support for CTI connections has improved over the last several releases, and each cluster can support a maximum of 40,000 CTI connections.
- The number of call processing subscribers in a cluster cannot exceed 4, plus 4 standby, for a total of 8 call processing subscriber nodes. Also, the total number of server nodes in a cluster, including the publisher, TFTP, and media servers, may not exceed 21 servers as the maximum allowed in a cluster.
- The name of a Unified CM virtual machine (VM) configuration corresponds to the maximum number of users, assuming that on average, each user has one phone. If this is not the case, the VM configurations would indicate the maximum number of endpoints registered to a Unified CM node. For example, a 10k-user VM configuration supports a maximum of 10,000 users, assuming one device per user. However, if you plan to deploy multiple devices per user, then the maximum number of supported users is reduced. For example, if you have 2 devices per user, then the 10k-user VM configuration would support a maximum of 5,000 users with 10,000 devices. This same principal applies for the smaller Unified CM VM configurations as well.

Deployment Options

The following deployment options are overall settings that affect all operations in the system, and they are independent of how many endpoints are registered or how many calls are in progress.

Database Complexity

The CPU usage is considerably higher when the configuration database in Unified CM is considered to be complex. There is no one metric to determine whether the database is simple or complex. As a general rule, the database is complex if you have configured more than a few thousand endpoints and more than a few hundred dial plan elements such as translation and route patterns, hunt pilots, and shared lines.

Number of Regions and Locations

Configuration of regions and locations in the Unified CM cluster requires both database and static memory. The number of gateways that can be defined in the cluster is also tied to the number of locations that can be defined. Table 25-3 lists these limits for some of the Unified CM VM configurations.

Table 25-3 Maximum Number of Regions, Locations, Gateways, and Trunks

VM Configuration	Maximum Number of Regions	Maximum Number of Locations	Maximum Number of Trunks and Gateways
1,000 or 2,500 Users	1,000	1,000	1,100
7,500 or 10,000 Users	2,000	2,000	2,100

Whether or not you can actually define the maximum number of locations and regions in a cluster depends on how "sparse" your codec matrix is. If you have too many non-default values in the inter-region codec setting, you might not be able to scale the system to its full capacity for regions and locations. As a general rule, the change from default should not exceed 10% of the maximum number.

Call Detail and Call Management Records

Generation of call detail records (CDR) and call management records (CMR) places a heavier burden on the CPU.

High Availability

After you determine the minimum number of nodes required for the specified deployment, add the desired number of additional subscriber nodes to provide redundancy. Redundancy options are described in the chapter on Call Processing, page 9-1.

Number of Virtual Server Nodes per Cluster

You can configure a regular cluster with up to four subscriber pairs. In a distributed topology, there may be multiple clusters even when none of the clusters has reached the maximum.

For a centralized topology, there is generally one cluster unless the capacity limit is reached. Note that other system limits might force a new cluster even if the per-node utilization is not at the limit.

Choice of VM Configurations and Hardware Platforms

Cisco provides Open Virtualization Archive (OVA) VM configurations that can be loaded onto a hypervisor. Different templates specify different capacities. For example, the 10,000 Users template defines a virtual machine that has a maximum capacity of 10,000 endpoints. There are also templates defined to support a maximum of 1,000, 2,500, and 7,500 endpoints.

The formal definitions of the VM configurations for Unified CM and other Unified Communications products are available at the following location:

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/collaboration-virtualization-sizing.html

Specific information for Unified CM is available at the following location:

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-cisco-unified-communications-manager.html

With Unified CM, some of the VM configurations are not supported on the low-end hardware platforms. To verify which VM configuration is supported on a hardware platform, refer to the documentation at:

http://www.cisco.com/go/virtualized-collaboration

Hardware and Virtualization Software Requirements

The following requirements are common to all applications. See each application's product documentation for additional requirements or restrictions.

- Details on supported and required virtualization hardware are available at:

  https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/collaboration-virtualization-hardware.html

- Details on supported and required virtualization software are available at:

  https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-software-requirements.html

Note Choice of placement of virtual machines running Unified CM and other Unified Communications products can have an impact on performance and availability. For a discussion of these and other considerations for Unified Communications on UCS deployments, refer to the documentation at http://www.cisco.com/go/virtualized-collaboration.

Endpoints

The number of endpoints is an important part of the overall load that the system must support. There are different types of endpoints, and each type imposes a different load on Unified CM. Endpoints can be differentiated by:

- Digital (IP) or analog (using an adaptor)
- Software-based or hardware
- The protocol supported (SIP or SCCP)
- Whether the endpoint is configured with security
- Dialing modes (en-bloc or overlap)
- Audio only or audio and video
- Other devices such as gateways (H.323 or MGCP)

Each endpoint configured in the system uses system resources (such as static memory) just by being defined and registered. The endpoint consumes CPU and dynamic memory based on its call rate.

An endpoint can also place additional load on the Unified CM by running applications such as CTI that interact with services running in the Unified CM.

Table 25-4 shows the maximum number of endpoints supported by different VM configuration types. Note that these values are guidelines only. A given system may support less than these maximum amounts because of other applications included in the deployment.

Table 25-4 Maximum Number of Endpoints Supported Per VM Configuration

VM Configuration	Maximum Endpoints per OVA Template[1]
10,000 Users	10,000
7,500 Users	7,500
2,500 Users	2,500
1,000 Users	1,000

1. These limits represent the maximum number of endpoints that can be configured in the database and registered per virtual subscriber node. All other registered devices such as media termination points (hardware of software) or SIP trunks do not count against these limits.

For Cisco Collaboration System Release (CSR) 12.*x*, the Unified CM deployments require all virtual nodes to increase their vRAM by 2 GB of memory for the following VM configuration templates:

- 1,000 users
  - 2 vCPU
  - 6 GB vRAM
  - 80 GB vDisk

- 2,500 users
  - 4 vCPU
  - 6 GB vRAM
  - 80 GB vDisk
- 7,500 users
  - 2 vCPU
  - 8 GB vRAM
  - 110 GB vDisk
- 10,000 users
  - 4 vCPU
  - 8 GB vRAM
  - 110 GB vDisk

For more details, refer to the documentation at:

http://www.cisco.com/go/virtualized-collaboration

Cisco Collaboration Clients and Applications

Cisco Collaboration Clients include the following software applications that run on user desktops or other access devices:

- Cisco Jabber Clients, page 25-18
- Cisco WebEx Connect, page 25-20
- Cisco UC Integration™ for Microsoft Lync, page 25-21
- Third-Party XMPP Clients and Applications, page 25-21

Cisco Jabber Desktop Client

Cisco Jabber provides the underlying services layer for several clients, including Cisco Jabber Clients for Windows and Mac and Cisco UC Integration™ for Microsoft Lync.

The Jabber Desktop Client provides two modes of operation, each of which uses different resources in Unified CM. When it operates in softphone mode, the Jabber Client acts as a SIP registered endpoint and contributes to the total number of endpoints in the system. When it operates in desk phone mode, the Jabber Client acts as a CTI agent and therefore uses CTI resources on Unified CM.

Users may switch the Jabber-based clients to work in either mode. Therefore, it is necessary to properly account for the system resources needed for the anticipated usage.

The following additional items must be considered for a Jabber Desktop Client deployment:

- Device Configuration

  When configured in softphone mode, a Jabber Desktop Client configuration file is downloaded through TFTP or HTTP to the client for Unified CM call control configuration information. In addition, any application dial rules or directory lookup rules are also downloaded through TFTP or HTTP to Jabber Desktop Client devices.

  The Jabber Desktop Client uses the Cisco Unified CM Cisco IP Phone (CCMCIP) service or UDS service to gather information about the devices associated with a user, and it uses this information to provide a list of IP phones available for control by the client in deskphone control mode. The Jabber Desktop Client in softphone mode uses the CCMCIP or UDS service to discover its device name for registration with Unified CM.

- Deskphone Mode

  When configured in deskphone mode, the Jabber Desktop Client establishes a CTI connection to Unified CM upon login and registration to allow for control of the IP phone. Unified CM supports up to 40,000 CTI connections. If you have a large number of clients operating in deskphone mode, make sure that you evenly distribute those CTI connections across all Unified CM subscribers running the CTIManager service. This can be achieved by creating multiple CTI Gateway profiles, each with a different pair of CTIManager addresses, and distributing the CTI Gateway profile assignments across all clients using deskphone mode.

- Voicemail

  When configured for voicemail, the Jabber Desktop Client updates and retrieves voicemail through an IMAP or REST connection to the mailstore.

- Authentication

  Client login and authentication, contact profile information, and incoming caller identification are all handled through a query to the LDAP directory, unless stored in the local Jabber Desktop Client cache.

- Contact Search

  There are several contact sources that can be used with the Jabber Desktop Client. For example, the UDS service can be used by clients to search for contacts in the Unified CM User database. Alternatively, LDAP integration can be used. If the requested contact cannot be found in the local Jabber Desktop Client cache, UDS or LDAP contact searches take place.

Cisco Jabber Clients

When designing and sizing a solution for Cisco Jabber Clients, you must consider the following scalability impacts for all the components:

- Client scalability

  The Cisco IM and Presence Service VM configuration template determines the number of users a cluster can support. The Cisco Jabber Client deployment must balance all users equally across all nodes in the cluster. This can be done automatically by setting the User Assignment Mode Sync Agent service parameter to **balanced**.

- IMAP scalability

  The number of IMAP or IMAP-Idle connections is determined by the messaging integration platform.

- Audio, video, and web conferencing

  Clients can access the conferencing services that are provided in your network. You need to account for these users when sizing the number of concurrent participants for these services. For additional information, refer to the chapter on Cisco Rich Media Conferencing, page 11-1.

Cisco Jabber Clients are supported on iPhone, iPad, and Android as mobile clients and on Windows and Mac as desktop clients. When sizing your deployment with Jabber Clients, keep in mind that users may have any combination of desktop and mobile clients. If the Multiple Device Messaging (MDM) feature is enabled for users, then each client that is associates to a user counts as a device and thus counts toward the total number of users supported in both the Unified CM and IM and Presence VM templates.

Note If a user has only a Jabber desktop client in desktop control mode, then that will count as only a single device due to the fact that the desk phone control utilizes CTI resources and lines.

The Cisco Jabber Clients interface with Unified CM. Therefore, the following guidelines for the current functionality of Unified CM apply when Cisco Jabber Client voice or video calls are initiated:

- CTI scalability

  In Desk Phone mode, calls from Cisco Jabber Clients use the CTI interface on Unified CM. Therefore, observe the CTI limits as defined in the chapter on Call Processing, page 9-1. You must include these CTI devices when sizing Unified CM clusters.

- Call admission control

  Cisco Jabber Client applies call admission control for voice and video calls by means of Unified CM locations or RSVP.

- Codec selection

  Cisco Jabber Client voice and video calls utilize codec selection through the Unified CM regions configurations.

- Cisco Unity Connection

  See the section on Managing Bandwidth, page 19-32, in the chapter on Cisco Voice Messaging, page 19-1.

- Cisco Jabber Clients are supported on iPhone, iPad, and Droid as mobile clients and on Windows PC and Mac as desktop clients.

  When sizing your deployment with Jabber Clients, keep in mind that users may have desktop and mobile clients, or multiple mobile clients or desktop clients. Users of the Multi Device Messaging (MDM) are more likely to request this feature. If it is enabled, then each client that is associates to a user counts as a device and hence counts against the total number of users supported by both the Unified CM and IM and Presence VM templates. If users have only Jabber desktop clients in desktop control mode, then they will count as only a single device due to the fact that the desk phone control utilizes CTI resources.

- Cisco WebEx Meetings Server

  Cisco WebEx Meetings Server provides WebEx conferencing services with voice, video, and collaboration sessions in a virtualized environment. For additional information about Cisco WebEx Meetings Server, refer to the *Cisco WebEx Meetings Server Planning Guide and System Requirements*, available at

  https://www.cisco.com/c/en/us/support/conferencing/webex-meetings-server/products-installation-and-configuration-guides-list.html

- Cisco Unified CM User Data Service (UDS)

  UDS is an umbrella of service APIs provided by Unified CM. UDS provides a contact source API that can be used by Jabber over Cisco Edge Series devices for contact source lookups. Using the UDS contact source to resolve contacts puts additional load on the system.

SAML SSO Cisco Jabber Client

Cisco Unified CM 10.*x* provides the Security Assertion Markup Language Single Sign-On (SAML SSO) feature, which enhances the end user experience by allowing users to log in only once to access all applications within the Cisco Collaboration solution.

SAML SSO provides secure mechanism to use credentials and relevant information of the end user to be leveraged across multiple Unified Communications applications (such as Unified CM, Cisco Unity Connection, and IM and Presence). For the SAML Single Sign-On feature to work as expected, the network architecture must scale to support the number of users for each cluster.

For a Unified Communications deployment across multiple applications (such as Unified CM, Cisco Unity Connection, and IM and Presence), all SAML requests must authenticate with the Identity Provider (IdP) for Cisco Jabber clients to login successfully.

Note SSO is supported by Unified Communications services with SAML.

Cisco Jabber with SAML SSO logins should also be factored into system sizing because the numbers of users logging into the system in a typical day at the same time could have an impact on the time it takes for user(s) to log in. This is expected due to the limiting factor of how many requests the system can process at one time. The current maximum login rate for Jabber users is 2.7 logins per second (about 166 logins per minute) or 10,000 logins within one hour. This is assuming that all users and devices are evenly distributed across all nodes and that Cisco Jabber is in softphone mode.

There are many interdependent variables that can affect Unified CM cluster scalability (such as regions, locations, gateways, media resources, and so forth). Therefore it is vital to determine the number of users, endpoints, and calls per user per hour, to deploy efficiently so that resources are available to handle the required load.

As an example, consider a deployment with redundant subscriber pairs supporting 5,000 users, each associated with two devices (desk phone and soft phone). This deployment would require the following number of virtual machines and VM configurations (assuming high availability and redundancy):

- One pair of Unified CM subscribers with 10k-user VM configurations
- One pair of IM and Presence 5k-user VM configurations

The IM and Presence 5k-user VM configuration pair would support the 5,000 users, and a pair of Unified CM 10k-user VM configurations would support the 10,000 devices.

Cisco WebEx Connect

A single end-user requires only a 56 kbps dial-up Internet connection to be able to log in to the Cisco WebEx Messenger service and get the basic capabilities such as presence, instant messaging, and VoIP calling. However, for a small office or branch office, a broadband connection with a minimum of 512 kbps is required in order to use the advanced features such as file transfer and screen capture.

For additional information on network and desktop requirements, refer to the latest version of the *Cisco WebEx Messenger Administration Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/webex-messenger/products-installation-guides-list.html

The Cisco Unified Communications integrations use Unified CM CTI Manager for click-to-call applications, as well as deskphone control mode with the Cisco Unified Client Services Framework. Therefore, observe the CTI limits as defined in the section on Applications and CTI, page 25-23. When Cisco UC Integration™ for Connect is operating in a softphone (audio on computer) mode, the Cisco Jabber Desktop Client is a SIP registered endpoint with Cisco Unified CM. When sizing a solution involving Cisco Unified Communications, you must include the CTI devices and the SIP endpoint devices utilizing resources on the Unified CM clusters.

Cisco UC Integration™ for Microsoft Lync

Cisco UC Integration™ for Microsoft Lync uses Unified CM CTI Manager for click-to-dial applications and deskphone control mode. Therefore, observe the CTI limits as defined in the chapter on Call Processing, page 9-1. When Cisco UC Integration™ for Microsoft Lync is operating in a softphone (audio on computer) mode, the client is a SIP registered endpoint with Cisco Unified CM. When sizing a solution involving Cisco Unified Communications, you must include the CTI devices and the SIP endpoint devices utilizing resources on the Unified CM clusters.

Third-Party XMPP Clients and Applications

Third-party Extensible Messaging and Presence Protocol (XMPP) clients may be used with both the WebEx Messenger service platform and the Cisco IM and Presence Service. Voice, video, and other collaboration mechanisms (except for instant messaging and chat) are typically not supported with these clients. Depending on their capabilities, these clients may be counted against the device capacities supported by the above products on their servers.

Mobile Unified Communications

Mobility in Unified Communications is multi-faceted. Each of the different aspects of mobile communications consumes different Unified CM resources and must be accounted for both independently and as a part of the whole system. The following sizing considerations apply to mobility, but note that aspects of mobility that do not affect Unified CM are not discussed here.

Cisco Unified Mobility

There are two parameters that are key to Unified CM's capacity to support single number reach (formerly Mobile Connect) and enterprise two-stage dialing (Mobile Voice Access and Enterprise Feature Access). For these functions to work appropriately, users must be enabled for mobility and remote destinations with shared lines must be defined for the users. Table 25-5 shows the limits for users and remote destinations and mobility identities in a cluster consisting of each class of Unified CM VM configurations.

Table 25-5 Maximum Number of Mobility Users and Remote Destinations and Mobility Identities per Cluster

Cluster Nodes	Maximum Number of Users Enabled for Mobility per Cluster	Maximum Number of Remote Destinations and Mobility Identities per Cluster
10,000 Users VM configuration	40,000	40,000 (or 10,000 per node)
7,500 Users VM configuration	30,000	30,000 (or 7,500 per node)
2,500 Users VM configuration	10,000	10,000 (or 2,500 per node)
1,000 Users VM configuration	4,000	4,000 (or 1,000 per node)

Note A mobility-enabled user is defined as a user that has a remote destination profile and at least one remote destination or a dual-mode device and a mobility identity configured.

Each remote destination and mobility identity defined in the system affects Unified CM in several ways:

- The remote destination or mobility identity occupies static memory and configuration space in the database.
- Each occurrence uses a shared line with the users primary device, and hence calls to that line use more CPU resources.
- If the remote destination or mobility identity is an external number (such as the user's cell phone or home), then gateway resources will be used to extend the call.

Call Traffic

The quantity and quality of call traffic is a very significant factor in sizing Unified CM.

It is important to differentiate between call types because call origination and termination are considered as distinct events in the half-call model. For endpoints registered on the same subscriber node, that subscriber handles both call halves for calls between these endpoints. For calls made between two subscriber nodes in the same cluster, each of the participating subscribers will handle either the call origination or call termination. For calls made between endpoints registered on different clusters, each cluster will handle only half of each call. For calls made between an endpoint in a cluster and the PSTN, a PSTN gateway will handle half of the call, and these call types form the basis for sizing the gateways.

For accurate sizing of call traffic, you must consider the following factors:

- Overall Busy Hour Call Attempts (BHCA) per user
- Average Call Holding Time (ACHT) per call
- BHCA from and to the PSTN using MGCP, H.323, and SIP protocols
- BHCA from and to other clusters using H.323 intercluster trunks or SIP protocols
- BHCA within the cluster

Each different type of call takes a different amount of CPU resources to set up. The number of busy hour call attempts determines the CPU usage. CPU requirements vary directly with the call placement rate. The ACHT determines the dynamic memory requirements to sustain calls for their duration. A longer ACHT means that more dynamic memory must remain allocated, thus increasing the memory requirement.

Call traffic can arise from other sources as well. Each time a call is redirected in a transfer or to voicemail, it requires processing by the CPU. If a directory number is configured on multiple phones, an incoming call to that number needs to be presented to all of those phones, thus increasing CPU usage at call setup time. If advanced features are being used, calls made using this technology, and the percentage of these calls that need to be redirected to the PSTN because of call quality, must also be accounted for.

Dial Plan

The dial plan in Unified CM consists of configuration elements that determine call routing and associated policies. In general, dial plan elements occupy static memory space in Unified CM. The following dial plan elements impact the amount of memory required:

- Directory numbers
- Shared directory numbers and the average number of endpoints that share the same DN
- Partitions, calling search spaces, translations, and transformation patterns
- Route patterns, route lists, and route groups
- Advertised and learned DN patterns
- Hunt pilots and hunt lists
- Circular, sequential, and broadcast line groups and their membership

There are no hard limits enforced by Unified CM for any of the dial plan elements, but there is a fixed amount of shared system memory available.

Most of the dial plan elements do not have a direct effect on CPU usage. The exception is shared lines, such as hunt lists and line groups. Each shared line multiplies the CPU cost of a call setup because the call is presented to all of the endpoints that share a particular directory number.

Applications and CTI

In the context of Unified CM, applications are the "extra" functions beyond simple call processing provided by Unified CM. In general these applications make use of Computer Telephone Integration (CTI), which allows users to initiate, terminate, reroute, or otherwise monitor and treat calls. Features such as Cisco Unified CM Assistant, Attendant Console, Contact Center, and others, depend on CTI to function.

Although the large Unified CM VM configurations are able to support CTI for all of their registered devices, the smaller VM configurations do not scale that high. Table 25-6 lists the maximum number of CTI resources supported for each Unified CM VM configuration. These maximum values apply to the following types of CTI resources:

- The maximum number of CTI controlled and/or monitored endpoints that can be registered to a Unified CM subscriber node.
- The maximum number of endpoints that a Unified CM subscriber node running the CTI Manager service can monitor or control.
- The maximum number of TAPI/JTAPI application instances that can connect to a Unified CM subscriber node running the CTI Manager service. The TAPI/JTAPI application instances that can connect to a Unified CM subscriber node running the CTI Manager service are sometimes referred as CTI connections.

Note that the maximum number of CTI resources for a VM configuration corresponds to the endpoint capacity of that VM configuration.

In addition to native applications provided by Unified CM, third-party applications may also be deployed that use Unified CM CTI resources. When counting CTI ports and route points, be sure to account for the third-party applications as well.

Table 25-6 CTI Resource Limits in Unified CM

VM Configuration	Maximum CTI Resources per Virtual Machine
1,000 Users	1,000
2,500 Users	2,500
7,500 Users	5,000 or 7,500[1]
10,000 Users	10,000

1. 7,500 CTI resources supported with Unified CM 10.5 and later releases; 5,000 CTI resources supported with Unified CM releases prior to 10.5.

In addition to the maximum number of connections and devices, CTI limits are also influenced by:

- The number of lines on each of the controlled devices (up to 5 lines per controlled device)
- The number of shared occurrences of a line controlled by CTI (up to 5 per line)
- The number of active CTI applications (up to 5 for any device)
- A maximum of 6 BHCA per controlled device

The CTI resources available on Unified CM are reduced if any of these values is exceeded.

Determining CTI Resources Required for a Unified CM Cluster

Use the following steps to determine the required number of CTI resources for a Unified CM cluster.

Step 1 Determine the total CTI device count.

Count the number of CTI devices that will be in use on the cluster.

Step 2 Determine the CTI line factor.

Determine the CTI line factor of all devices in the cluster, according to Table 25-7.

Table 25-7 CTI Line Factor

Number of Lines per CTI Device	CTI Line Factor
1 to 5 lines	1.0
6 lines	1.2
7 lines	1.4
8 lines	1.6
9 lines	1.8
10 lines	2.0

Note If there are multiple line factors for the devices within a cluster; determine the average line factor across all CTI devices in the system.

Step 3 Determine the application factor.

Determine the application factor of all devices in the cluster, according to Table 25-8.

Table 25-8 CTI Application Factor

Number of Applications per CTI Device	CTI Application Factor
1 to 5 applications	1.0
6 applications	1.2
7 applications	1.4
8 applications	1.6
9 applications	1.8
10 applications	2.0

Step 4 Calculate the required number of CTI resources according to the following formula:

Required Number of CTI Resources = (Total CTI Device Count) * (The greater of {the CTI Line Factor or the CTI Application Factor})

The following examples illustrate the process.

Example 1: 500 CTI devices deployed with an average of 9 lines per device and an average of 4 applications per device. According to the factor lists in Table 25-7 and Table 25-8, the 9 lines per device renders a line factor of 1.8, while 4 applications per device renders an application factor of 1.0. Applying these values in the formula from Step 4 yields:

(500 CTI Devices) * (Greater of {1.8 Line Factor or 1.0 Application Factor})

(500 CTI Devices) * (1.8 Line Factor) = 900 total CTI resources required

Example 2: 2,000 CTI devices deployed with an average of 5 lines per device and an average of 9 applications per device. According to the factor lists in Table 25-7 and Table 25-8, the 5 lines per device renders a line factor of 1.0, while 9 applications per device renders an application factor of 1.8. Applying these values in the formula from Step 4 yields:

(2000 CTI Devices) * (Greater of {1.0 Line Factor or 1.8 Application Factor})

(2000 CTI Devices) * (1.8 Application Factor) = 3,600 total CTI resources required

Example 3: 5,000 CTI devices deployed with an average of 2 lines per device and an average of 3 applications per device. According to the factor lists in Table 25-7 and Table 25-8, the 2 lines per device renders a line factor of 1, while 3 applications per device renders an application factor of 1. Applying these values in the formula from Step 4 yields:

(5,000 CTI Devices) * (Greater of {1 Line Factor or 1 Application Factor})

(5,000 CTI Devices) * (1 Line or Application Factor) = 5,000 total CTI resources required

IP Phone Services

Cisco Unified IP Phone Services are applications that utilize the web client and/or server and XML capabilities of the Cisco Unified IP Phone. The Cisco Unified IP Phone firmware contains a micro-browser that enables limited web browsing capability. These phone service applications provide the potential for value-added services and productivity enhancement by running directly on the user's desktop phone.

Cisco Unified IP Phone Services act, for the most part, as HTTP clients. In most cases they use Unified CM only as a redirect server to the location of the subscribed service. Because Unified CM acts only as a redirect server, there typically is minimal performance impact on Unified CM unless there is a large number of requests (hundreds of requests per minute or more).

With the exception of IP Phone Services for the integrated Extension Mobility and Unified CM Assistant applications, IP Phone Services must reside on a separate web server. Running phone services other than Extension Mobility and Unified CM Assistant on a Unified CM node is not supported.

Cisco Extension Mobility and Extension Mobility Cross Cluster

Using Extension Mobility (EM) impacts the system performance in the following ways:

- Creation of EM profiles requires both disk database space and static memory.
- The rate at which users may log into their EM accounts affects both CPU and memory usage. Unified CM nodes have bounds on the maximum number of logins per minute that they can support.
- Extension Mobility Cross Cluster (EMCC) has a higher impact on resources. There is a limit on the number of EMCC users that a Unified CM node can support. The maximum EMCC login rates supported are lower than those supported for EM. In addition, there is a trade-off between EM and EMCC login rates. If both are occurring at the same time, then the maximum capacity for each will be reduced.
- EM and EMCC login rates per cluster are not simply the login rate of each node multiplied by the number of nodes in the cluster, because profiles in a shared database have to be accessed. The maximum login rate in a cluster consisting of more than one call processing subscriber should be limited to 1.5 times that of a single node.

Table 25-9 shows the maximum number of EM and EMCC logins per minute for each type of VM configuration.

Table 25-9 EM and EMCC Rates Per VM Configuration

VM Configuration	Maximum EM Login Rate (per Node)	Maximum EM Login Rate (Dual Nodes)	Maximum EMCC Login Rate (Per Node)	Maximum EMCC Login Rate (Dual Nodes)	Maximum Concurrent EMCC Devices
1,000 Users	200	300	60	70	333
2,500 Users	235	352	71	80	833
7,500 or 10,000 Users	250	375	75	90	2,500

Cisco Extension Mobility login and logout functionality can be distributed across a pair of subscriber nodes to increase login/logout cluster capacity. For example, when the EM load is distributed evenly between two virtual machines with the 7,500-user VM configuration, the maximum cluster-wide capacity is 375 sequential logins and/or logouts per minute.

Note The Cisco Extension Mobility service can be activated on more than two nodes for redundancy purposes, but Cisco supports a maximum of two subscriber nodes actively handling logins/logouts at any given time.

Note Enabling EM Security does not diminish performance.

The EMCC login/logout process requires more processing resources than intracluster EM login/logout, therefore the maximum supported login/logout rates are lower for EMCC. In the absence of any intracluster EM logins/logouts, Unified CM supports a maximum rate of 75 EMCC logins/logouts per minute with a virtual machine using the 7,500-user or 10,000-user VM configuration. Most deployments will have a combination of intracluster and intercluster logins/logouts occurring. For this more common scenario, the mix of EMCC logins/logouts (whether acting as home cluster or visiting cluster) should be modeled for 40 per minute, while the intracluster EM logins should be modeled for 185 logins/logouts when using a single EM server node. The intracluster EM login rate can be increased to 280 logins/logouts per minute when using the 7,500-user or 10,000-user VM configuration in a dual EM node configuration. (See Table 25-9.)

EMCC logged-in devices (visiting phones) consume twice as many resources as any other endpoint in a cluster. The maximum supported number of EMCC logged-in devices is 2,500 per cluster, but this also decreases the theoretical maximum number of other devices per cluster from 30,000 to 25,000. Even if the number of other registered devices in the cluster is reduced, the maximum supported number of EMCC logged-in devices is still 2,500.

Cisco Unified CM Assistant

The Cisco Unified CM Assistant application uses CTI resources in Unified CM for line monitoring and phone control. Each line (including intercom lines) on a Unified CM Assistant or Manager phone requires a CTI line from the CTIManager. In addition, each Unified CM Assistant route point requires a CTI line instance from the CTIManager. When you configure Unified CM Assistant, the number of required CTI lines or connections must be considered with regard to the overall cluster limit for CTI lines or connections.

The following limits apply to Unified CM Assistant:

- A maximum of 10 Assistants can be configured per Manager.
- A maximum of 33 Managers can be configured for a single Assistant (if each Manager has one Unified CM Assistant-controlled line).
- A maximum of 3,500 Assistants and 3,500 Managers (7,000 total users) can be configured per cluster using the 7,500-user or 10,000-user virtual machines
- A maximum of three pairs of primary and backup Unified CM Assistant nodes can be deployed per cluster if the **Enable Multiple Active Mode** advanced service parameter is set to **True** and a second and third pool of Unified CM Assistant server nodes are configured.

In order to achieve the maximum Unified CM Assistant user capacity of 3,500 Managers and 3,500 Assistants (7,000 users total), multiple Unified CM Assistant server pools must be defined. (For more information, see Unified CM Assistant, page 18-19.)

Cisco WebDialer

Cisco WebDialer provides a convenient way for users to initiate a call. Its impact on Unified CM is fairly limited because extra resources are required only at call initiation and are not tied up for the duration of the call. Once the call has been established, its impact on Unified CM is just like any other call.

The WebDialer and Redirector services can run on one or more subscriber nodes within a Unified CM cluster, and they support the following capacities:

- Each WebDialer service can handle up to 4 call requests per second per node.
- Each Redirector service can handle up to 8 call requests per second.

The following general formula can be used to determine the number of WebDialer calls per second (cps):

(Number of WebDialer users) * ((Average BHCA) / (3600 seconds/hour))

When performing this calculation, it is important to estimate properly the number of BHCA per user that will be initiated specifically from using the WebDialer service. The following example illustrates the use of these WebDialer design calculations for a sample organization.

Example: Calculating WebDialer Calls per Second

Company XYZ wishes to enable click-to-call applications using the WebDialer service, and their preliminary traffic analysis resulted in the following information:

- 10,000 users will be enabled for click-to-call functionality.
- Each user averages 6 BHCA.
- 50% of all calls are dialed outbound, and 50% are received inbound.
- Projections estimate 30% of all outbound calls will be initiated using the WebDialer service.

Note These values are just examples used to illustrate a WebDialer deployment sizing exercise. User dialing characteristics vary widely from organization to organization.

10,000 users each with 6 BHCA equates to a total of 60,000 BHCA. However, WebDialer deployment sizing calculations must account for placed calls only. Given the initial information for this sizing example, we know that 50% of the total BHCA is for placed or outbound calls. This results in a total of 30,000 placed BHCA for all the users enabled for click-to-call using WebDialer.

Of these placed calls, the percentage that will be initiated using the WebDialer service will vary from organization to organization. For the organization in this example, several click-to-call applications are made available to the users, and it is projected that 30% of all placed calls will be initiated using WebDialer.

(30,000 placed BHCA) * 0.30 = 9,000 placed BHCA using WebDialer

To determine the number of WebDialer server nodes required to support a load of 9,000 BHCA, we convert this value to the average call attempts per second required to sustain this busy hour:

(9,000 call attempts / hour) * (hour/3,600 seconds) = 2.5 cps

Each WebDialer service can support up to 4 cps, therefore one node can be configured to run the WebDialer service in this example. This would allow for future growth of WebDialer usage. In order to maintain WebDialer capacity during a server node failure, additional backup WebDialer server nodes should be deployed to provide redundancy.

Attendant Console

The integration of Cisco Unified CM with the Attendant Console utilizes CTI resources. The server-based attendant console monitors the last 2,000 users to whom the attendant sent calls, thus increasing CTI resource usage. In addition, each call uses a number of CTI route points and ports for greetings, queuing, and so forth.

Media Resources

Unified CM offers the Cisco IP Voice Media Streaming Application (IPVMS), which provides certain media functions that are performed in software only and do not require hardware resources. Unified CM can act as a media termination point (MTP), as a conference bridge, as an annunciator (for playing announcements), or as a source of music-on-hold streams. Although the capabilities of Unified CM are limited compared to similar functions provided by Cisco Integrated Service Routers (ISRs), they are generally the key source of music-on-hold streams (both unicast and multicast).

The Cisco IP Voice Media Streaming Application may be deployed in one of two ways:

- Co-resident deployment

  In a co-resident deployment, the streaming application runs on any server node (either publisher or subscriber) in the cluster that is also running the Unified CM software.

  Note The term *co-resident* refers to two or more services or applications running on the same server node or virtual machine.

- Standalone deployment

  In a standalone deployment, the streaming application runs on a dedicated server node within the Unified CM cluster. The Cisco IP Voice Media Streaming Application service is the only service enabled on the server node, and the only function of the server node is to provide media resources to devices within the network.

The Cisco IP Voice Media Streaming Application can provide MTP, annunciation, and conferencing capabilities, but a more scalable design is to place these functions on external Cisco Integrated Service Routers (ISRs). The music-on-hold functionality of this application is, however, not so easily placed on external sources. Table 25-10 lists the maximum values that may be configured for each of these services.

Table 25-10 Cisco IP Voice Media Streaming Application Capacity Limits

Media Device Type	Default Quantity	Maximum Number of Streams or Devices	Supported Codecs
Annunciator	48	750	G.711, G.729, L16WB
Software Conference Bridge	48	256	G.711, L16WB
Music on Hold	250	1,000	G.711, G.729, L16WB
Software Media Termination Point (MTP)	48	512	G.711, L16WB, passthrough

The following notes apply to Table 25-10:

- All values represent the number of callers supported per media device. For instance, 48 software conference bridges can support 16 three-party conferences.
- These devices can be co-resident with the call processing nodes when using default settings or near to default settings.
- When increasing capacities to the maximum values, Cisco recommends deploying the media devices on standalone nodes (not with call processing).
- If MoH audio sources are used with initial (greeting) announcements, Cisco recommends keeping the initial announcements less than 15 seconds in duration, otherwise you might need to reduce the maximum number of MoH streams per MoH server node to between 500 and 700 due to extra file I/O.
- Each media device may be disabled/enabled via the IPVMS Service Parameter (MoH is on the MoH device configuration page). It is possible to configure an MoH-only Unified CM node, and so forth.

Note To calculate the capacities of each of the media functions on the DSPs supported by each individual ISR, refer to the Cisco ISR product data sheets or to the chapter on Media Resources, page 7-1.

Music on Hold

Table 25-11 11 lists the VM configurations and the maximum number of simultaneous music-on-hold (MoH) streams each node can support. You should ensure that the actual usage does not exceed these limits, because once MoH maximum stream capacity has been reached, additional load could result in poor MoH quality, erratic MoH operation, or even loss of MoH functionality. Add additional MoH nodes (co-resident or dedicated) to increase Unified CM cluster MoH stream capacity.

Table 25-11 Music on Hold Maximum Per-Node Stream Capacity

Unified CM OVA Template	Unified CM 10.5(2) and Later		Unified CM 10.5(1) and Earlier	
	Co-resident MoH Streams (non-sRTP)[1]	Standalone MoH Streams	Co-resident MoH Streams	Standalone MoH Streams
1,000 User	500	750	500	500
2,500 User			1,000	1,000
7,500 User	750	1,000	1,000	1,000
10,000 User				

1. All capacities based on non-sRTP streams.

As shown in Table 25-12, beginning with Unified CM 10.5(2) you can define a maximum of 500 unique sources of audio for Music on Hold in a Unified CM cluster. The maximum audio source capacities shown in Table 25-12 are per-cluster based on the VM configuration size and MoH server type (co-resident or standalone) used in the cluster. Adding MoH nodes to a Unified CM cluster increases only MoH stream capacity but does not increase audio source capacity. Audio source capacity can be increased only by moving from co-resident to standalone MoH nodes, increasing the cluster-wide node VM configuration size, or adding additional Unified CM clusters.

Table 25-12 Music on Hold Maximum Per-Cluster Audio Source Capacity

Unified CM OVA Template	Unified CM 10.5(2) and Later		Unified CM 10.5(1) and Earlier	
	Co-resident MoH Sources	Standalone MoH Sources	Co-resident MoH Sources	Standalone MoH Sources
1,000 User	100	250	50	
2,500 User				
7,500 User	250	500		
10,000 User				

The capacity limits described in Table 25-11 and Table 25-12apply to any combination of unicast, multicast, or simultaneous unicast and multicast streams.

Performance Considerations

To maximize the number of MoH audio sources and streams, you must reduce the number of some other media devices, such as disabling software MTPs and/or software conference bridges. The Cisco IP Voice Media Streaming Application service does not support maximum settings for all the media devices simultaneously. Oversubscribing the system resources (for example, CPU usage and disk I/O) with media devices would impact the overall system performance. An IPVMS alarm is issued if a media device is unable to meet provisioned capacity.

For low-end configurations (1,000-user or 2,500-user VM configuration) and MoH co-resident with moderate call processing, MoH is limited to a maximum of 500 streams, 100 MoH audio sources, and 48 to 64 annunciator streams with MTPs and conference bridges set at default values or disabled.

A dedicated 1,000-user or 2,500-user VM configuration MoH node is required to support 750 MoH streams with 250 MoH audio sources and 250 annunciator streams.

To support a maximum of 1,000 MoH streams, 500 MoH audio sources, and 750 annunciators, the minimum requirement is a 7,500-user OVA dedicated standalone MoH server.

Use of sRTP for MoH and/or annunciator will reduce the maximum number of MoH callers by 25%, and a dedicated IPVMS server for MoH and annunciator is highly recommended in this case.

The Unified CM MoH server supports four codecs: G.711 ulaw, G.711 mulaw, G729a, and Wideband audio. With unicast MoH, because the codec is negotiated during call setup, the number of MoH streams depends not on the number of MoH codecs enabled but on the number of endpoints that are on hold with unicast MoH. In the case of multicast MoH, each multicast-enabled audio source generates one MoH steam for each MoH codec enabled. For example, if 2 codecs are enabled and all 500 MoH sources are multicast-enabled, then 1,000 multicast MoH streams would be active even if no endpoints are on hold. In this scenario, if any endpoints are placed on unicast MoH, then additional MoH streams capacity would be required.

Impact on Unified CM

Whether deployed in co-resident or standalone mode, the Cisco IP Voice Media Streaming Application consumes CPU and memory resources. This impact must be considered in the overall sizing of Unified CM.

In general, usage of media resources can be considered to add to the BHCA that needs to be processed by Unified CM.

Call Queuing (Hunt Pilot Queuing)

The maximum number of media streams that can be sent for call queuing is the same as with Music on Hold streams. See Music on Hold, page 25-30, for details.

The maximum number of hunt pilots with call queuing enabled is 100 per Unified CM subscriber node. The maximum number of simultaneous callers in queue for each hunt pilot is 100. The maximum number of members across all hunt lists does not change when call queuing is enabled.

LDAP Directory Integration

The Unified CM Database Synchronization feature provides a mechanism for importing a subset of the user configuration data (attributes) from the LDAP store into the Unified CM publisher database. Once synchronization of a user account has occurred, the copy of each user's LDAP account information may then be associated to additional data required to enable specific Unified Communications features for that user. When authentication is also enabled, the user's credentials are used to bind to the LDAP store for password verification. The end user's password is never stored in the Unified CM database when enabled for synchronization and/or authentication.

User account information is cluster-specific. Each Unified CM publisher node maintains a unique list of those users receiving Unified Communications services from that cluster. Synchronization agreements are cluster-specific, and each publisher has its own unique copy of user account information.

The maximum number of users for a Unified CM cluster is limited by the maximum size of the internal configuration database that gets replicated between the cluster members. Currently the maximum number of users that can be configured or synchronized is 160,000. To optimize directory synchronization performance, Cisco recommends considering the following points:

- Directory lookup from phones and web pages may use the Unified CM database or the IP Phone Service SDK. When directory lookup functionality uses the Unified CM database, only users who were configured or synchronized from the LDAP store are shown in the directory. If a subset of users is synchronized, then only that subset of users is seen on directory lookup.
- When the IP Phone Services SDK is used for directory lookup, but authentication of Unified CM users to LDAP is needed, the synchronization can be limited to the subset of users who would log in to the Unified CM cluster.
- If only one cluster exists, if the LDAP store contains fewer than the maximum number of users supported by the Unified CM cluster, and if directory lookup is implemented to the Unified CM database, then it is possible to import the entire LDAP directory.
- If multiple clusters exist and if the number of users in LDAP is less than the maximum number of users supported by the Unified CM cluster, it is possible to import all users into every cluster to ensure directory lookup has all the entries.
- If the number of user accounts in LDAP exceeds the maximum number of users supported by the Unified CM cluster and if the entire user set should be visible to all users, it will be necessary to use the Unified IP Phone Services SDK to off-load the directory lookup from Unified CM.
- If both synchronization and authentication are enabled, user accounts that have either been configured or synchronized into the Unified CM database will be able to log in to that cluster. The decision about which users to synchronize will impact the decision on directory lookup support.

Note Cisco supports the synchronization of user accounts up to the limit mentioned above, but it does not enforce this limit. Synchronizing more user accounts can lead to starvation of disk space, slower database performance, and longer upgrade times.

Cisco Unified CM Megacluster Deployment

A Unified CM cluster is considered to be a megacluster when the number of call processing subscribers exceeds the normal cluster maximum of 4 pairs. A megacluster may have up to 8 pairs of call processing subscribers and no more than 21 server nodes in a single megacluster.

For example, you may have the publisher, TFTP, TFTP backup, MoH, MoH backup, 8 primary, and 8 backup servers counted toward the 21-server limit.

Note IM and Presence does not count toward the 21-server limit for a megacluster deployment.

Cisco IM and Presence has introduced a VM configuration template to align with megacluster deployments using a 25,000-user VM configuration.

A Unified Communications deployment can be simplified in certain cases with a Unified CM megacluster. The following limits increase with such a deployment:

- Maximum number of endpoints supported is twice the number of a normal cluster (8 call processing subscriber pairs).
- Maximum number of CTI devices and connections also doubles.

However, some cluster-wide constants do not increase. Chief among these are:

- Size of the configuration database
- Number of locations and regions
- Maximum number of LDAP synchronized or provisioned end users (160,000 users per cluster)

Note Due to the many potential complexities surrounding megacluster deployments, customers who wish to pursue such a deployment must engage their Cisco Account Team, Cisco Advanced Services, or their certified Cisco Unified Communications Partner.

Cisco IM and Presence

As with all other applications, sizing for Cisco IM and Presence is accomplished in the following way:

- Decompose the system into its most elemental services.
- Measure the unit cost of each of these services.
- Analyze the given system description as an aggregation of the identified services and arrive at a net system cost.
- Determine the number of required servers based on system cost and deployment options.

For IM and Presence, the following system variables in the system under analysis are relevant and must be considered for accurate sizing:

- Number and type of users
  - Clients employed by users to obtain presence services
  - Operating mode for users (instant messaging only or full Unified Communications facilities)
- Presence-related activities performed by typical users
  - Contact list size and composition (intracluster, intercluster, and federated). The Cisco IM and Presence system architecture is based on an average contact list size of 75 contacts per user on a fully populated system. While per-user contact list size will vary across the system, if significant numbers of users on the system exceed the average list size of 75 contacts, system performance will be impacted. By default the maximum contact list size is 200. If some users will exceed 200 contacts, this maximum contact list size can be changed by modifying the Presence Settings of the IM and Presence cluster.
  - Number of instant messages (directly between two users) per user during the busy hour
  - Chat support with number of chat rooms, users per chat room, and instant messages per user per chat room
  - State changes per user (both call related and user initiated)
- Deployment model
  - Whether intercluster presence is supported
  - Whether federation is supported
  - Whether high availability is desired
- Server preferences
  - The desired VM configuration size

- System options
  - Whether compliance recording is required

Once the system requirements are quantified, the number of required virtual machines can be determined from the data in Table 25-13.

Table 25-13 Maximum Number of Users Supported per IM and Presence Cluster[1]

VM Configuration	Maximum Users Supported in Full Unified Communications Mode
500 Users	1,500
1,000 Users	1,000
2,000 Users	6,000
5,000 Users	15,000
15,000 Users	45,000
25,000 Users	75,000

1. Maximum supported sub-clusters is 3.

Roster Management

The number of contacts and watchers a user has, does impact the system performance. Due to the potential severity of the impact, the system administrator must monitor the usage to ensure that the cluster average per user does not exceed 75 contacts and/or watchers.

By default the service parameters are set to a maximum of 200 contacts and 200 watchers per user. The intent of this default parameter setting is to provide options for users who require a higher number of contacts. This does not imply that all 15,000 presence users on an IM and Presence node may each have 200 contacts and watchers.

We recommend that all IM and Presence deployments do not exceed a cluster average of 75 contacts and/or watchers per user, even though the service parameter is set to 200 for both.

For example, assume that we have the 15,000 Users VM configuration template in a fully loaded cluster with 3 sub-clusters and 45,000 presence-enabled users. If we want to maintain an average of 75 contacts for every user in the cluster, then the maximum number of contacts allowed for the entire cluster would be:

(45,000 users) * (75 contacts/user) = 3.375M contacts allowed for the IM and Presence cluster

Some users in this cluster may have up to 200 contacts while other users have fewer contacts, as long as the total number of contacts for all users in the cluster does not exceed 3.375M.

As another example, assume that we have a deployment of 5,000 IM and Presence users, and 50 of those users need 1,000 contacts each. The maximum number of contacts allowed for this deployment would be:

(5,000 users) * (75 contacts/user) = 375,000 contacts allowed for the entire deployment

The 50 heavy users would need: (50 users) * (1,000 contacts/user) = 50,000 contacts. That would leave (375,000 – 50,000) = 325,000 contacts available for the remaining 4,950 users, or:

325,000/4,950 = approximately 65 contacts on average for each of the other 4,950 users

For additional information on Cisco IM and Presence, refer to the latest version of the *Compatibility Matrix for Cisco Unified Communications Manager and the IM and Presence Service*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-device-support-tables-list.html

The formal definitions of the VM configurations for Cisco IM and Presence are available at

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-cisco-ucm-im-presence.html

Impact on Unified CM

The Cisco IM and Presence Service influences the performance of Unified CM in the following ways:

- User synchronization through an AXL/SOAP interface
- Presence information through a SIP trunk
- CTI traffic to enable phone control

In general, the impact of user synchronization (except for a one-time hit) and that of presence information through the SIP trunk are negligible. The effect of CTI control of phones, however, must be counted against CTI limits.

IM and Presence VM configurations differ from Unified CM VM configurations. IM and Presence templates are user based while Unified CM templates are device based. For example, a 5k-user IM and Presence VM configuration used with a Unified CM 10k-user VM configuration would support 5,000 users with 2 devices each. All IM and Presence nodes within the same cluster must use the same type of VM configuration.

Note Prior to IM and Presence release 11.5, concurrent user logins were limited to a maximum of 80% of the IM and Presence VM template capacity. With IM and Presence 11.5 and later releases, 100% of the presence users can log in through Jabber at the same time. For example, in a deployment of 45,000 presence-enabled users, IM and Presence releases prior to 11.5 support only 36,000 (80% of 45,000) concurrent logins, while IM and Presence 11.5 and later releases support all 45,000 users logged in at the same time (assuming only one Jabber client per user login). This enhancement also increases the allowed number of concurrent Jabber users by 20%.

Centralized IM and Presence

Cisco IM and Presence supports a centralized deployment option. A centralized IM and Presence cluster can provide presence service for users on multiple remote Unified CM clusters; however, the total number of users across all the remote Unified CM clusters must not exceed 75,000, assuming that each user has a single client. Multiple clients per user would reduce this limit.

Note The centralized IM and Presence cluster requires a Unified CM publisher node, for a total of 7 servers in the cluster: 3 IM and Presence sub-cluster pairs (6 servers) + the Unified CM publisher node.

For deploying a centralized IM and Presence cluster, we recommend using the 25k-user IM and Presence VM template for all the IM and Presence nodes in the cluster and using the 10k-user Unified CM VM template for the Unified CM publisher node of that centralized cluster.

The centralized IM and Presence deployment can be clustered over the WAN, subject to the following restrictions:

- All remote Unified CM clusters must be within 80 ms round-trip-time (RTT) of the centralized IM and Presence cluster.
- A centralized IM and Presence cluster may be connected to another centralized IM and Presence cluster by means of an intercluster trunk with a maximum latency of 300 ms RTT.

Emergency Services

The Cisco Emergency Responder tracks the locations of phones and the access switch ports to which they are connected. The phones may be discovered automatically or entered manually into the Emergency Responder. Table 25-14 shows the VM configurations that support the Emergency Responder and their maximum capacities.

Note These limits apply to standalone Emergency Responder deployments, and they assume that Native Emergency Services are not being used.

Table 25-14 Cisco Emergency Responder VM Configurations and Capacities

VM Configuration	Maximum Number of Automatically Tracked Phones	Maximum Number of Manually Configured Phones	Maximum Number of Roaming Phones	Maximum Number of Switches	Maximum Number of Switch Ports	Maximum Number of Emergency Response Locations
12,000 Users	12,000	2,500	1,200	500	30,000	3,000
20,000 Users	20,000	5,000	2,000	1,000	60,000	7,500
30,000 Users	30,000	10,000	3,000	2,000	120,000	10,000
40,000 Users	40,000	12,500	4,000	2,500	150,000	12,500

The formal definitions of the VM configurations for Cisco Emergency Responder and other Unified Communication products are available at the following location:

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-cisco-emergency-responder.html

There can be only one Emergency Responder active per Unified CM cluster. Therefore, choose an VM configuration that has sufficient resources to provide emergency coverage for all of the phones in the cluster.

For more details on network hardware and software requirements for Emergency Responder, refer to the *Cisco Emergency Responder Administration Guide*, available at

https://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

Cisco Expressway

Cisco Expressway deployments rely on Cisco Unified CM as the component for call control, including remote endpoint registration. When sizing such a system, consider the function it performs as well as its impact to Unified CM.

When sizing Cisco Expressway, you typically must consider the following parameters to determine the required number of Cisco Expressway-C and Expressway-E node pairs:

- Number of endpoint registrations through each pair of Expressway-C and Expressway-E nodes during peak usage time
- Expected number of simultaneous voice-only and video calls traversing each pair of Expressway-C and Expressway-E nodes

Expressway-C and Expressway-E clusters support a maximum of 6 nodes.

Mobile and remote access does not require any specific licenses, but business-to-business communication requires rich media licenses. Licenses in the form of rich media sessions are shared across an Expressway cluster. Each Expressway node in the cluster contributes its assigned rich media sessions to the cluster database, which is then shared across all of the nodes in the cluster. This model results in any one Expressway node being able to carry many more licenses than its physical capacity.

Cisco Expressway Capacity Planning

Table 25-15 lists the Cisco Expressway proxy registrations and call capacities for Cisco Expressway-C and Expressway-E server node pairs and clusters.

Table 25-15 Cisco Expressway-C and Expressway-E Node and Cluster Capacities

Platform	Proxy Registrations[1]	Video Calls	Audio-only Calls
Large OVA (or Expressway Appliance)	2,500 per node 10,000 per cluster	500 per node 2,000 per cluster	1,000 per node 4,000 per cluster
Medium OVA (or Expressway Appliance)	2,500 per node 10,000 per cluster	100 per node 400 per cluster	200 per node 800 per cluster
Small OVA (Business Edition 6000)	2,500 per node 2,500 per cluster[2]	100 per node 100 per cluster[2]	200 per node 200 per cluster[2]

1. Proxy registration applies only to mobile and remote access connections, not business-to-business communications.
2. Cisco Expressway-C and Expressway-E can be clustered across multiple Business Edition 6000 nodes for redundancy purposes; however, there is no increased capacity when clustering with Business Edition 6000.

Note The large OVA template is supported only with limited hardware. Refer to the documentation at https://www.cisco.com/go/virtualized-collaboration for more information.

The following guidelines apply when clustering Cisco Expressway:

- Expressway clusters support up to 6 nodes (cluster capacity up to 4 times the node capacity).
- The capacity of all nodes across and within each Expressway-E and Expressway-C cluster pair must be the same. For example, an Expressway-E node using the large VM configuration must not be deployed if other nodes in the Expressway-E cluster or in the corresponding Expressway-C cluster are using the medium size VM configuration.

- Expressway peers should be deployed in equal numbers across Expressway-E and Expressway-C clusters. For example, a three-node Expressway-E cluster should be deployed with a three-node Expressway-C cluster.
- An Expressway-E and Expressway-C cluster pair can be formed by a combination of nodes running on an appliance or running as a virtual machine, as long as the node capacity is the same across all nodes.
- The Expressway node VM configurations or Expressway Appliances must match across and within Expressway Series cluster pairs.
- Multiple pairs of Expressway Series clusters may be deployed to increase capacity.

Note There is a dependency between Cisco Expressway clusters and Cisco Unified CM clusters. Expressway capacity planning must also consider the capacity of the associated or dependent Unified CM cluster(s).

For more information about Cisco Expressway capacity planning considerations, including sizing limits, capacity planning, and deployment considerations, refer to the Cisco Expressway product documentation available at

https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/tsd-products-support-series-home.html

Gateways

PSTN gateways handle traffic between the Unified Communications system and the PSTN. The amount of traffic determines the resource usage (CPU and memory) and the number of PSTN DS0 circuits required for the gateways.

PSTN traffic is generated by the endpoints registered to Unified CM, but there may be other sources such as interactive voice response (IVR) applications and other parts of a contact center deployment.

Gateways can also perform other functions that require resources (such as CPU, memory, and DSP). These functions include media processing such as media termination point (MTP), transcoding, conference bridge, and RSVP Agents.

Gateways, especially those based on the Cisco Integrated Service Routers (ISRs), can provide other functions such as serving as VXML processing engines, acting as border elements, doubling as Cisco Unified Communications Manager Express or Survivable Remote Site Telephony (SRST), or performing WAN edge functions. All of these activities need to be taken into account when calculating the gateway load.

Gateway Groups

When considering the number of gateways, you also need to consider the geographical placement of physical gateway servers. In a deployment model where PSTN access is distributed, you need to size those gateways as a group by themselves and assign the appropriate amount of load to each such group.

A grouping might also be appropriate if certain gateways are expected to be dedicated for certain functions and share common characteristics.

Therefore, to accurately estimate the number of gateways required, the following information is required:

- Groups of gateways that share a common group profile. The common profiles will depend on the complexity of the deployment.

- For each group, the traffic patterns, platform, blocking probability, and so forth, that make up the profile.
- The individual gateway platform that makes up the group. In deciding on a particular gateway model, ensure that the model can support the capabilities and the capacity that is expected of it. Note that more than one gateway might be required in a gateway group, depending on the ability of the selected platform to meet the performance requirements.

PSTN Traffic

PSTN circuits are shared by all users of the system, and there are usually many more users than PSTN circuits. The number of circuits required is estimated by using the traffic management principles described in the section on call traffic (Call Traffic, page 25-22).

The amount of external traffic received and generated by your business determines the number of PSTN circuits required. When converting from a TDM-based system, many customers will continue to use the same number of circuits for their IP-based communications system as they had used for the previous system. However, you may want to perform a new traffic analysis, which will detect if the system is over-provisioned for the current levels of traffic (and, therefore, the customer is paying for circuits that are not needed). If the system is under-provisioned, users will experience an unacceptable number of blocked and/or lost calls, in which case increasing the number of circuits will remedy the situation.

The number of PSTN circuits determines the DSP requirements for the gateways. DSP resources are required to perform conversion between IP and TDM voice (PSTN circuits use TDM encoding).

One key input is the blocking factor, which determines the percentage of call attempts that may not be serviced at peak traffic levels. A lower blocking factor means that more call attempts will succeed, but the system will require more circuits than for a higher blocking factor.

Gateway Sizing for Contact Center Traffic

Short call durations as well as bursty call arrival rates impact the PSTN gateway's ability to process the traffic. Under these circumstances the gateway needs more resources to process all calls in a timely manner, compared to calls of longer duration that are presented more uniformly over time. Because gateways have varying capabilities to deal with these traffic patterns, careful consideration should be given to selecting the appropriate gateway for the environment in which it will operate. Some gateways support more T1/E1 ports than others, and some are more able than others to deal with multiple calls arriving at the same time.

For a traffic pattern with multiple calls arriving in close proximity to each other (that is, high or bursty call arrival rates), a gateway with a suitable rating of calls per second (cps) is the best fit. Under these conditions, for example, the Cisco 3945 Integrated Services Router can maintain 28 cps with 420 calls active at once.

For traffic patterns with a steady arrival rate, the maximum number of active calls that a gateway can handle is generally the more important consideration. Under these conditions, using calls with 180-second hold times, for example, the Cisco 3945 Integrated Services Router can maintain 720 simultaneously active calls with a call arrival rate of up to 4 cps.

These numbers assume that all of the following conditions apply:

- CPU utilization does not exceed 75%.
- PSTN gateway calls are made with ISDN PRI trunks using H.323.
- The Real Time Control Protocol (RTCP) timer is set to the default value of 5 seconds.
- Voice Activity Detection (VAD) is off.

- G.711 uses 20 ms packetization.
- Cisco IOS Release 15.0.1M is used.
- Dedicated voice gateway configurations are used, with Ethernet (or Gigabit Ethernet) egress and no QoS features. (Using QoS-enabled egress interfaces or non-Ethernet egress interfaces, or both, will consume additional CPU resources.)
- No supplementary call features or services are enabled – such as general security (for example, access control lists or firewalls), voice-specific security (TLS, IPSec and/or SRTP), AAA lookups, gatekeeper-assisted call setups, VoiceXML or TCL-enabled call flows, call admission control (RSVP), and SNMP polling/logging. Such extra call features use additional CPU resources.

Voice Activity Detection (VAD)

Voice Activity Detection (VAD) is a digital signal processing feature that suppresses the creation of most of the IP packets during times when the speech path in a particular direction of the call is perceived to be silent. Typically only one party on a call speaks at a time, so that packets need to flow in only one direction, and packets in the reverse (or silent) direction need not be sent except as an occasional keepalive measure. VAD can therefore provide significant savings in the number of IP packets sent for a VoIP call, and thereby save considerable CPU cycles on the gateway platform. While the actual packet savings that VAD can provide varies with the call flow, the application, and the nature of speaker interactions, it tends to use 10% to 30% fewer packets than would be sent for a call made with VAD turned off.

VAD is most often turned off in endpoints and voice gateways deployed in Unified CM networks; VAD is most often turned on in voice gateways in other types of network deployments.

Codec

Both G.711 and G.729A use as their default configuration a 20 ms sampling time, which results in a 50 packets-per-second (pps) VoIP call in each direction. While a G.711 IP packet (200 bytes) is larger than a G.729A packet (60 bytes), this difference has not proven to have any significant effect on voice gateway CPU performance. Both G.711 and G.729 packets qualify as "small" IP packets to the router, therefore the packet rate is the salient codec parameter affecting CPU performance.

Performance Overload

Cisco IOS is designed to have some amount of CPU left over during peak processing, to handle interrupt-level events. The performance figures in this section are measured with the processor running at an average load of approximately 75%. If the load on a given Cisco IOS gateway continually exceeds this threshold, the following results will occur:

- The deployment will not be supported by Cisco Technical Assistance Center (TAC).
- The Cisco IOS Gateway will display anomalous behavior, including Q.921 time-outs, longer post-dial delay, and potentially interface flaps.

Cisco IOS Gateways are designed to handle a short burst of calls, but continual overloading of the recommended call rate (calls per second) is not supported.

Note With any gateway, you might be tempted to assign unused hardware ports to other tasks, such as on a Cisco Communication Media Module (CMM) gateway where traffic calculations have dictated that only a portion of the ports can be used for PSTN traffic. However, the remaining ports must remain unused, otherwise the CPU will be driven beyond supported levels.

Performance Tuning

The CPU utilization of a Cisco IOS Voice Gateway is affected by every process that is enabled in a chassis. Some of the lowest level processes such as IP routing and memory defragmentation will occur even when there is no live traffic on the chassis.

Lowering the CPU utilization can help to increase the performance of a Cisco IOS Voice Gateway by ensuring that there are enough available CPU resources to process the real-time voice packets and the call setup instructions. Table 25-16 describes some of the techniques for decreasing CPU utilization.

Table 25-16 Techniques for Reducing Gateway CPU Utilization

Technique	CPU Savings	Description
Enable Voice Activity Detection (VAD)	Up to 20%	Enabling VAD can result in up to 45% fewer voice packets in typical conversations. The difficultly is that, in scenarios where voice recognition is used or there are long delays, a reduction in voice quality can occur. Voice appears to "pop" in at the beginning and "pop" out at the end of talk spurts.
Disable Real Time Control Protocol (RTCP)	Up to 5%	Disabling RTCP results in less out-of-band information being sent between the originating and terminating gateways. This results in lower quality of statistics displayed on the paired gateway. This can also result in the terminating gateway having a call "hang" for a longer period of time if RTCP packets are being used to determine if a call is no longer active.
Disable other non-essential functions such as: Authentication, Authorization, and Accounting (AAA); Simple Network Management Protocol (SNMP); and logging	Up to 2%	Any of these processes, when not required, can be disabled and will result in lower CPU utilization by freeing up the CPU to provide faster processing of real-time traffic.
Change the call pattern to increase the length of the call (and reduce the number of calls per second)	Varies	This can be done by a variety of techniques such as including a long(er) introduction prompt played at the beginning of a call or adjusting the call script at the call center.

Additional Information

A full discussion of every gateway, its capabilities, and call processing capacities is not possible in this chapter. For more information on Cisco Voice Gateways, refer to the following documentation:

- Cisco Voice Gateway Solutions:

  https://www.cisco.com/c/en/us/products/unified-communications/communications-gateways/index.html

- Interfaces and signaling types supported by the following Cisco Voice Gateways:
  - Cisco 3900 Series Integrated Services Routers

    https://www.cisco.com/c/en/us/products/routers/3900-series-integrated-services-routers-isr/relevant-interfaces-and-modules.html
  - Cisco 2900 Series Integrated Services Routers

    https://www.cisco.com/c/en/us/products/routers/2900-series-integrated-services-routers-isr/relevant-interfaces-and-modules.html

- Gateway features supported with MGCP, SIP, and H.323:

  https://www.cisco.com/c/dam/en/us/products/collateral/routers/2800-series-integrated-services-routers-isr/product_data_sheet0900aecd8057f2e0.pdf

- SIP gateway RFC compliance:

  https://www.cisco.com/c/en/us/products/collateral/unified-communications/ios-gateways-session-initiation-protocol-sip/product_data_sheet0900aecd804110a2.html

- Skinny Client Control Protocol (SCCP) feature support with FXS gateways:

  https://www.cisco.com/c/en/us/products/collateral/unified-communications/vg-series-gateways/product_data_sheet09186a00801d87f6.html

- Gateway capacities and minimum releases of Cisco IOS and Unified CM required for conferencing, transcoding, media termination point (MTP), MGCP, SIP, and H.323 gateway features:

  https://www.cisco.com/c/dam/en/us/products/collateral/routers/2800-series-integrated-services-routers-isr/product_data_sheet0900aecd8057f2e0.pdf

Voice Messaging

Voice messaging is an application that needs to be sized not only by itself but also for its effect on other Unified Communications components, mainly Unified CM.

Total number of users is the key factor for sizing the voice messaging system. Other factors that affect sizing for voice messaging are:

- Number of calls during the busy hour that the application has to handle
- Average length of messages left on the servers
- Number of users who check their messages during the busy hour
- Average length of user sessions
- Any advanced operations such as voice recognition or text-to-speech sessions

- Any media transcoding
- Ports on the voice messaging system are analogous to the DS0s on a gateway and are shared resources that need to be optimized. The same considerations of probabilistic arrival and the need for blocking apply to both types of resources.

Table 25-17 shows the applicability of the various voice messaging solutions to the scalability requirements of the deployment.

Table 25-17 Scaling Voice Messaging Solutions

Solutions	Maximum Number of Users Supported on a Single Node (or Failover or Clustered Deployment)				Maximum Number of Users Supported in a Digital Networking Solution	Maximum Number of Users Supported in an HTTPS Networking Solution
	500	1,000	15,000	20,000	100,000	100,000
Cisco Unity Express	Yes	No	No	No	Yes	No
Cisco Business Edition	Yes	Yes	No	No	No	No
Cisco Unity Connection (Unified/Integrated Messaging and Cisco Business Edition 7000)	Yes	Yes	Yes	Yes	Yes	Yes

Table 25-18 shows the maximum limits of various functions of different VM configurations running Cisco Unity Connection.

Table 25-18 VM Configurations and Capacities for Cisco Unity Connection

VM Configuration	Maximum Number of Ports	Maximum Voice Recognition Sessions	Maximum Text to Speech Sessions	Maximum Number of Voicemail Users
100 Users	8	8	8	100
500 Users	16	16	16	500
1,000 Users	24	24	24	1,000
5,000 Users	100	100	100	5,000
10,000 Users	150	150	150	10,000
20,000 Users	250	250	250	20,000

The formal definitions of the VM configurations for Cisco Unity Connection are available at

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-cisco-unity-connection.html

Impact on Unified CM

The impact of a voice messaging system on Unified CM can be gauged by considering the extra processing that Unified CM needs to do. These extra call flows add to the sizing load of Unified CM as follows:

- Calls that need to be forwarded to the voice messaging system when the user is not present or if the user deliberately forwards the calls using Do Not Disturb (DND) or other features.
- Calls from users who dial the voice messaging pilot number to access their voice messages go through Unified CM, and these calls must be added to the calls being handled by Unified CM, including both the number and the duration of these calls.

Collaborative Conferencing

Cisco Collaborative Conferencing systems include Cisco Unified CM as a component for call control. When sizing such a system, the function it performs as well as its impact to Unified CM should be considered.

When sizing such conferencing systems, you typically have to consider the following parameters to determine the type and number of nodes:

- Number of users who could use the system at any one time
- Number of audio, video, and web users on the system at the peak usage time
- Required dial-in duration
- Video resolution and audio codec requirements

Sizing Guidelines for Audio Conferencing

Cisco recommends the following methods for calculating audio conferencing capacity:

- Calculation based on average monthly usage

  If you know the average voice conferencing usage (average minutes per month), use Table 25-19 to calculate the audio conferencing capacity.

Table 25-19 Audio Conferencing Capacity Based on Average Monthly Usage

Average Monthly Usage (minutes)	Baseline Usage (minutes per port per month)	Estimated Number of Ports
20,000 to 50,000	1,500	15 to 35
50,000 to 500,000	2,000	25 to 250
500,000 to 1,000,000	3,000	165 to 335
1,000,000 to 2,000,000	3,500	285 to 570
2,000,000 to 8,000,000	4,000	500 to 2,000

- Calculation based on number of users

  You should plan on having one port for every 20 users with average usage. If the users are heavy conference users, then provision one port for every 15 users. For example, in a system with 6000 users, you should provision 300 audio ports; however, if those users heavily use conferencing, then plan for 400 audio ports.

- Calculation based on actual peak usage

  Actual voice conferencing usage during peak hours usually can be obtained from existing voice conferencing system logs or service provider bills. Cisco recommends provisioning 30% extra capacity based on the actual peak usage in order to protect against extra conferencing volume.

Factors Affecting System Sizing

In addition to the estimates provided by the methods described above for the system baseline port requirement, the following factors also affect system sizing:

- When migrating from an "operator-scheduled" model to a user-scheduled model, you might need to add another 20% to the baseline.
- The default average meeting size is 4.5 callers per meeting. Use the value that is applicable to your case if it is different than the default.
- Increase the baseline estimate accordingly if the following condition applies:

  (Estimated meetings per day) * (Estimated users) > 80% of baseline
- If the largest single meeting exceeds 20% of the estimated capacity, increase the estimate accordingly.
- If there are continuous meetings with dedicated ports, then you must add those additional ports ((Meetings) * (Dedicated callers)) to the baseline.

The total number of ports will include all the above factors in addition to the baseline. Plan for conferencing system capacity expansion if the total estimated port capacity exceeds 80% of the maximum supported ports.

Sizing Guidelines for Video Conferencing

Cisco recommends the following three methods for calculating video conferencing capacity:

- Calculation based on number of knowledgeable workers

  Cisco recommends provisioning a video user license for every 40 knowledgeable workers.
- Calculation based on number of voice conferencing user licenses

  Cisco recommends provisioning video conferencing capacity in the range of 17% to 25% of existing audio user licenses. The percentage depends on business requirements regarding video conferencing and on the size of the conferencing system.
- Calculation based on existing video Multipoint Control Unit (MCU)

  Cisco recommends deploying a direct replacement for an existing video conferencing system.

Impact on Unified CM

The impact to Unified CM can be analyzed based on the extra call traffic that the conferencing system generates. The most impact occurs when conference users dial into their meetings that are typically scheduled at the top of the hour or half-hour. A large amount of call traffic within a few minutes of conference start times increases the load on Unified CM for just those few minutes and must be designed in appropriately. In addition, if conference users include callers from the PSTN or from other clusters, those parameters must also be considered to gauge their impact on the gateways.

Cisco WebEx Meetings Server

The Cisco WebEx Meetings Server provides WebEx conferencing services using enterprise-provided servers (a Cisco UCS server clusters in the enterprise data center).

Cisco WebEx Meetings Server is offered in different configurations, which the sizing tool chooses based primarily on the number of knowledge workers that have access to the conferencing service.

For each configuration, Cisco recommends a standard Cisco UCS server type with specific configurations of hardware and VMware products. However, Cisco WebEx Meetings Server is designed to work on any equivalent or better Cisco UCS Server that meets or exceeds these specifications.

This product is packaged as a VMware vSphere compatible OVA virtual appliance and not as a collection of software packages on a DVD. Cisco WebEx Meetings Server requires the vCenter product to deploy the OVA and install the Cisco WebEx Meetings Server product.

Currently, Cisco WebEx Meetings Server does not operate in co-resident mode on the Cisco UCS server. Cisco WebEx Meetings Server requires a dedicated UCS server.

For additional information about Cisco WebEx Meetings Server, refer to latest version of the *Cisco WebEx Meetings Server Planning Guide and System Requirements*, available at

https://www.cisco.com/c/en/us/support/conferencing/webex-meetings-server/products-installation-and-configuration-guides-list.html

Sizing Factors

The sizing tool uses the following inputs to calculate system capacity:

- Number of knowledge users

  The number of knowledge users is defined as the set of employees that can access the conferencing system (to initiate a conference or join a conference).

  Many knowledge users share the available conferencing ports. The assumption is that only a small percentage of users are active in a conference call at any time. Based on this percentage, we can estimate of the number of conferencing ports required to support these users.

  The sizing tool defines light usage (3.3% of users active at any one time), average usage (5% active) and heavy usage (10% active). Therefore, a system operating with average usage will support twice as many users as a system with heavy usage.

- User minutes per month

  The user minutes per month is the total number of minutes of active conferences for the month, across all ports. This value is expressed in thousands of minutes. This factor is significant for calculating the size of the recording server.

- Actual peak usage

  Actual peak usage is defined as the maximum number of concurrent users of the system. This number is significant in determining the required number of conferencing ports. Cisco recommends provisioning enough capacity to handle 30% more users than the actual peak usage, to ensure that adequate conferencing ports are available during peak usage times.

- Video

  The percent of conferences with video and high-quality video will impact the network bandwidth required by the system. Up to 50% of the users can be using high-quality video.

- Traffic mix

  Different call types require different Unified CM resources. For accurate assessment of the Unified CM impact, the tool requires estimates of the following call types:

  - Percent of conference calls incoming via enterprise IP phones. This call leg is handled by Unified CM and therefore has an impact on Unified CM capacity.
  - Percent of external call legs, which impacts sizing for PSTN gateways.

- Access by external users

  If external users need to access the system, additional virtual machines are configured to provide reverse proxy functionality. If the system is intended for internal users only, these additional virtual machines are not required.

- Disaster recovery

  For disaster recover, you can configure a cold-standby system in a second data center. If the primary system is configured for high availability, you can optionally choose to configure high availability for the disaster recovery system.

- High availability

  The system can be configured in non-redundant mode or in high-availability (HA) mode. In HA mode, the cluster is provisioned with one or more backup servers (the specific configuration depends on the system size).

System Capacities

Cisco WebEx Meetings Server is offered in four system sizes, as listed in Table 25-20. System size is expressed as the maximum number of concurrent users of the system. Maximum concurrent users defines the maximum number of users who can participate in conference calls at any given time.

Table 25-20 VM Configurations and Capacities for Cisco WebEx Meeting Server

Maximum	50 Concurrent Users	250 Concurrent Users	800 Concurrent Users	2,000 Concurrent Users
Audio and web users (combined)	50	250	800	2,000
Video and video sharing (combined)	25	125	400	1,000
Participants in a single meeting	50	100	100	100
Playback recordings of meetings that have ended	12	63	200	500
Recordings of meetings in progress	3	13	40	100
Number of conferences (average of 2 participants per meeting)	25	125	400	1,000
Calls per second	1	3	8	20

Note that the following optional capabilities can be used without any impact on system capacity:

- Encrypted audio (sRTP)
- Secured Meeting Center Web (SSL)
- Different audio codecs
- Low-resolution video

Recordings

Meetings for up to 5% of the ports (or 10% of meetings) can be recorded. You need to provision an NFS-mounted hard drive of sufficient size to store the recorded meetings. One meeting will generate a file with a size of 50 to 100 MB.

Network Bandwidth

To estimate the bandwidth required on the LAN and WAN, the sizing tool makes the following assumptions:

- Each port will use 1 Mbps of network bandwidth.
- The user mix will be 80% internal to the enterprise and 20% external.

Therefore, the required bandwidth (in Mbps) on the LAN is 0.8 * (Number of ports), and on the WAN is 0.2 * (Number of ports)

Cisco Prime Collaboration Management Tools

Cisco Prime Collaboration offers a set of integrated tools to test, deploy, and monitor Cisco Unified Communications and TelePresence systems. Cisco Prime Collaboration includes the following products: Prime Collaboration Provisioning, Prime Collaboration Assurance, and Prime Collaboration Analytics.

These applications run on virtual machines. Cisco Prime Collaboration Provisioning runs on its own virtual machine, while Cisco Prime Collaboration Assurance and Cisco Prime Analytics run on the same virtual machine. Virtual machine sizing for these applications is relatively simple and depends directly on the number of endpoints or network devices that they are expected to manage.

Cisco Prime Collaboration Provisioning

Cisco Prime Collaboration Provisioning can support up to 150,000 endpoints and is implemented either on a single machine (for up to 10,000 endpoints) or on two machines (over 10,000 endpoints).

Virtual machine resources required for various levels of performance are described in the latest version of the Cisco Prime Collaboration install and upgrade guides, available at

https://www.cisco.com/c/en/us/support/cloud-systems-management/prime-collaboration/products-installation-guides-list.html

Cisco Prime Collaboration Assurance

Cisco Prime Collaboration Assurance can manage phones and other network devices such as routers and switches. It operates in a single machine configuration and supports up to 150,000 phones.

Virtual machine resources required for various levels of performance are described in the latest version of the *Cisco Prime Collaboration Quick Start Guide*, available at

https://www.cisco.com/c/en/us/support/cloud-systems-management/prime-collaboration/products-installation-guides-list.html

Cisco Prime Collaboration Analytics

Cisco Prime Collaboration Analytics runs on the same virtual machine as Cisco Prime Collaboration Assurance and works with Cisco Network Analysis Modules (NAMs) to measure voice quality.

Hardware resources required for various levels of performance are described in the latest version of the *Cisco Prime Collaboration Data Sheet*, available at

https://www.cisco.com/c/en/us/products/cloud-systems-management/prime-collaboration/datasheet-listing.html

Sizing for Standalone Products

The following products are not included in the sizing tools, but the following sections describe how to size these products:

- Cisco Unified Communications Manager Express, page 25-49
- Cisco Business Edition, page 25-49

Cisco Unified Communications Manager Express

Cisco Unified Communications Manager Express (Unified CME) runs on one of the Cisco IOS Integrated Services Router (ISR) platforms, from the low-end Cisco 881 ISR to the high-end Cisco 3945E ISR 2. Each of these routers has an upper limit on the number of phones that it can support. The actual capacity of these platforms to do call processing may be limited by the other functions that they perform, such as IP routing, Domain Name System (DNS), Dynamic Host Control Protocol (DHCP), and so forth.

Unified CME can support a maximum of 450 endpoints on a single Cisco IOS platform; however, each router platform has a different endpoint capacity based on the size of the system. Because Unified CME is not supported within the Cisco Collaboration Sizing Tool, it is imperative to follow the capacity information provided in the Unified CME product data sheets available at

https://www.cisco.com/c/en/us/products/unified-communications/unified-communications-manager-express/datasheet-listing.html

Cisco Business Edition

Cisco Business Edition is a packaged collaboration solution that is preloaded with premium services for voice, video, mobility, messaging, conferencing, instant messaging and presence, and contact center applications.

The Cisco Business Edition 4000 (BE4000) is the newest addition to the Business Edition Family. The BE4000 is powered by Cisco Unified Communication Manager Express and provides call processing services for small to medium single-site deployments and deployments in which local call processing at a remote site is needed.

The BE4000 is a dedicated cloud-managed platform that provides audio telephony and voicemail service for up to 200 audio telephony devices, with each device licensed for telephony and a voicemail port.

The BE4000 supports a maximum of 200 users with the following:

- Cisco IP Phone 7800 Series and 8800 Series SIP endpoints
- Cisco Unity Express Virtual Voicemail
- Maximum of 5 busy hour call attempts (BHCA)

Cisco Business Edition 6000 and 7000 both have platform model options to choose from.

Cisco Business Edition 6000 is available in three hardware platform options:

- BE6000H — Maximum capacity of 1,000 users; 2,500 devices; and 100 contact center agents. Supports nine collaboration application options in a single virtualized server platform. Maximum of 5,000 BHCA.
- BE6000M — Maximum capacity of 1,000 users; 1,200 devices; and 100 contact center agents. Supports five collaboration application options in a single virtualized server platform. Maximum of 5,000 BHCA
- BE6000S — Maximum capacity of 150 users and 300 devices. Supports five fixed collaboration applications in a single integrated router/gateway/virtualized blade server platform. Maximum of 750 BHCA.

To learn more about Cisco Business Edition 6000 solutions, visit https://www.cisco.com/go/be6000.

Cisco Business Edition 7000 is available in two hardware platform options:

- BE7000H — This high-density model typically supports five to ten collaboration applications in deployments sized for 1,000 to 5,000 users with 3,000 to 15,000 devices and multiple sites.
- BE7000M — This medium-density model typically supports four to six collaboration applications in deployments sized for 1,000 to 5,000 users with 3,000 to 15,000 devices and multiple sites.

To learn more about Cisco Business Edition 7000 solutions, visit https://www.cisco.com/go/be7000.

Busy Hour Call Attempts (BHCA) for Cisco Business Edition

This section use Cisco Business Edition 6000H as an example to calculate capacity, but the information in this section also applies to BE6000M as well as the smaller 750 BHCA capacity BE6000S.

As mentioned above, Business Edition 6000H supports a maximum of 5,000 BHCA. When calculating your system usage, stay at or below this BHCA maximum to avoid oversubscribing Cisco Business Edition 6000. The BHCA consideration becomes significant when the usage for any phone is above 4 BHCA. A true BHCA value can be determined only by taking a baseline measurement of usage for the phone during the busy hour. Extra care is needed when estimating this usage without a baseline.

Device Calculations for Cisco Business Edition 6000H

Devices can be grouped into two main categories for the purpose of this calculation: phone devices and trunk devices.

A phone device is a single callable endpoint. It can be any single client device such as a Cisco Unified IP Phone 8800 Series or other Collaboration voice and video endpoints, a software client such as Cisco Jabber, an analog phone port, or an H.323 client. While Cisco Business Edition 6000 supports a maximum of 300 endpoints on a BE6000S, 1,200 endpoints on a medium-density server, or 2,500 endpoints on a high-density server, as indicated above, actual endpoint capacity depends on the total system BHCA.

A trunk device carries multiple calls to more than one endpoint. It can be any trunk or gateway device such as a SIP trunk or a gatekeeper-controlled H.323 trunk. Business Edition 6000 supports intercluster trunking as well as H.323, SIP, and MGCP trunks or gateways and analog gateways. Cisco recommends using SIP trunks rather than the other protocols.

The method for calculating BHCA is much the same for both types of devices, but trunk devices typically have a much higher BHCA because a larger group of endpoints is using them to access an external group of users (PSTN or other PBX extensions).

You can define groups of devices (phone devices or trunk devices) with usage characteristics based on BHCA, and then you can add the BHCA for each device group to get the total BHCA for the system, always ensuring that you are within the supported maximum of 5,000 BHCA.

For example, you can calculate the total BHCA for 100 phones at 4 BHCA each and 80 phones at 12 BHCA each as follows:

100 phones at 4 BHCA is 100*4 = 400

80 phones at 12 BHCA is 80*12 = 960

Total BHCA = (100*4) + (80*12) = 1,360 BHCA for all phones

For trunk devices, you can calculate the BHCA on the trunks if you know the percentage of calls made by the devices that are originating or terminating on the PSTN. For this example, if 50% of all device calls originate or terminate at the PSTN, then the net effect that the device BHCA (1360 in this case) would have on the gateways would be 50% of 1360, or 680 BHCA. Therefore, the total system BHCA for phone devices and trunk devices in this example would be:

Total system BHCA = 1,360 + 680 = 2,040 BHCA

If you have shared lines across multiple phones, the BHCA should include one call leg (there are two call legs per each call) for each phone that shares that line. Shared lines across multiple groups of devices will affect the BHCA for that group. That is, one call to a shared line is calculated as one call leg per line instance, or half (0.5) of a call. If you have different groups of phones that generate different BHCAs, use the following method to calculate the BHCA value:

Shared line BHCA = 0.5 * (Number of shared lines) * (BHCA per line)

For example, assume there are two classes of users with the following characteristics:

100 phones at 8 BHCA = 800 BHCA

150 phones at 4 BHCA = 600 BHCA

Also assume 10 shared lines for each group, which would add the following BHCA values:

10 shared lines in the group at 8 BHCA = 0.5*10*8 = 40 BHCA

10 shared lines in the group at 4 BHCA = 0.5*10*4 = 20 BHCA

The total BHCA for all phone devices in this case is the sum of the BHCA for each phone group added to the sum of the BHCA for the shared lines:

800 + 600 + 40 + 20 = 1,460 total BHCA

Note that the total BHCA in each example above is acceptable because it is below the system maximum of 5,000 BHCA.

If you are using Cisco Unified Mobility for single number reach (SNR) on Business Edition 6000, keep in mind that calls extended to remote destinations and mobility identities or off-system phone numbers affect BHCA. In order to avoid oversubscribing the appliance, you have to account for this SNR remote destination or off-system phone BHCA. To calculate the BHCA for these SNR features, see Capacity Planning for Cisco Unified Mobility, page 21-74, and add that value to your total BHCA calculation.

Note Media authentication and encryption using Secure RTP (SRTP) impacts the system resources and affects system performance. If you plan to use media authentication or encryption, keep this fact in mind and make the appropriate adjustments. Typically, 100 IP phones without security enabled results in the same system resource impact as 90 IP phones with security enabled (10:9 ratio).

Another aspect of capacity planning to consider for Cisco Business Edition 6000 is call coverage. Special groups of devices can be created to handle incoming calls for a certain service according to different rules (top-down, circular hunt, longest idle, or broadcast). This is done through hunt or line group configuration within Cisco Business Edition 6000. BHCA can also be affected by this factor, but only as it pertains to the line group distribution broadcast algorithm (ring all members). For Business Edition 6000, Cisco recommends configuring no more than three members of a hunt or line group when a broadcast distribution algorithm is required. Depending on the load of the system, doing so could greatly affect the BHCA of the system and possibly oversubscribe the platform's resources. The number of hunt or line groups that have a distribution algorithm of broadcast should also be limited to no more than three. These are best practice recommendations meant to prevent over-subscription of the system BHCA. Exceeding these recommendations within a deployment is supported as long as the overall BHCA capacity of the system is not exceeded.

Mixing different types of hardware platforms within a Unified CM cluster is also allowed. However, because not all VM configurations are supported on all server platforms, mixing VM configurations will impact the overall cluster capacity, as described in the section on Mixing Hardware Platforms and Business Edition Platforms, page 9-8.

Cisco Unified Mobility for Cisco Business Edition 6000

The capacity for Cisco Unified Mobility users on Cisco Business Edition 6000 systems depends exclusively on both the number of remote destinations per user and the BHCA of the users enabled for Unified Mobility, rather than on server hardware. Thus, the number of remote destinations supported on Cisco Business Edition 6000 depends directly on the BHCA of these users.

Each configured remote destination or mobility identity has potential BHCA implications. For every remote destination or mobility identity configured for a user, one additional call leg is used. Because each call consists of two call legs, one remote destination ring is equal to half (0.5) of a call. Therefore, you can use the following formula to calculate the total remote destination BHCA:

Total remote destination and mobility identity BHCA = 0.5 * (Number of users) * (Number of remote destinations and mobility identities per user) * (User BHCA)

For example:

Assuming a system of 300 users at 5 BHCA each, with each user having one remote destination or mobility identity (total of 300 remote destinations and mobility identities), the calculation for the total remote destination and mobility identity BHCA would be:

Total remote destination and mobility identity BHCA =
0.5 * (300 users) * (1 remote destination or mobility identity per user) * (5 BHCA per user) =
750 BHCA

Total user BHCA in this example is [(300 users) * (5 BHCA per user)], which is 1,500 total user BHCA. By adding the total remote destination BHCA of 750 to this value, we get a total system BHCA of 2,250 (1,500 total user BHCA + 750 total remote destination and mobility identity BHCA).

If other applications or additional BHCA variables are in use on the system in the example above, the capacity might be limited. (See the preceding sections for further details.)

For more information on Cisco Business Edition 6000 capacity planning as well as other product information, refer to the following product documentation for Cisco Business Edition 6000:

- https://www.cisco.com/go/be6000
- https://www.cisco.com/c/en/us/support/unified-communications/business-edition-6000/tsd-products-support-series-home.html

Cisco Collaboration System Migration

Revised: March 1, 2018

This chapter provides recommendations for administrators to manage migrations from traditional PBX systems to IP telephony as well as from previous Cisco Collaboration System Releases (9.*x*, 10.*x*, and 11.*x*) to the latest Cisco Collaboration System Release (CSR) 12.*x*.

For more information on minimum hardware and software requirements for Open Virtualization Archive (OVA) templates, VMware, ESXi Hypervisor, and Collaboration applications, refer to the following documentation:

- *Virtualization Software Requirements*

  https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-software-requirements.html

- *VMware Compatibility Guide*

  https://www.vmware.com/resources/compatibility

For Cisco Collaboration System Release 12.*x*, most Cisco Collaboration applications require virtualization deployments and may not be installed directly on a server without a hypervisor. VMware vSphere ESXi is currently the only supported hypervisor, and it is mandatory for all virtualized deployments of Cisco Collaboration Systems. Cisco Collaboration System Releases do not support VMware vSphere ESX or any other VMware server virtualization products besides ESXi.

This chapter discusses the following types of migrations:

- Cisco 7800 Series MCS servers to Cisco Unified Communications Manager (Unified CM) on Cisco Unified Computing System (UCS) servers
- Cisco Video Communication Server (VCS) endpoint registration to Unified CM registration
- H.323 gateways and trunks to SIP gateways and trunks
- SCCP endpoints to SIP endpoints
- Numeric dialing to URI dialing
- Migration to Cisco Smart Software licensing and licensing management with Cisco Smart Software Manager (Cisco SSM)

In order to ensure a successful migration, Cisco recommends using the following resources to validate that all requirements have been met prior to migration:

- *Cisco Collaboration Systems Release Compatibility Matrix*

  https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/uc_system/unified/communications/system/Compatibility/CSR-Compatibility-Matrix.html

- *Compatibility Matrix for Cisco Unified Communications Manager and the IM and Presence Service*

  https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-device-support-tables-list.html

The validation will ensure a successful migration using supported upgrade paths. For example, some early application software versions might require a multi-step upgrade in order to migrate successfully. Similarly, server hardware along with software compatibility might require a combination of multi-step hardware and software upgrades.

For details on Cisco Collaboration System products, refer to the documentation available at

https://www.cisco.com/c/en/us/solutions/collaboration/collaboration-systems-release/index.html

For a list of all supported system hardware, refer to the Unified Computing Products documentation available at

https://www.cisco.com/c/en/us/products/servers-unified-computing/product-listing.html

What's New in This Chapter

Table 26-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 26-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Cisco Prime License Manager has been replaced by Cisco Smart Software Licensing	Cisco Smart Software Manager, page 26-9	March 1, 2018
Other minor updates for Cisco Collaboration System Release (CSR) 12.*x*.	Various sections of this chapter	March 1, 2018

Coexistence or Migration of Solutions

This is an important choice that must be made. Coexistence typically means two or more systems coexisting for an extended period of time (for example anything greater than six months.) Under this scenario feature transparency, whether for PBX, voicemail, or other features, becomes a more significant consideration. Investment and/or upgrades to existing systems might be necessary in order to deliver the level of feature transparency required. Migration typically occurs over a shorter period of time (for example, less than six months). Under this scenario users are more likely to tolerate a subset of existing features, knowing that the migration will be complete in a "short" time. Often existing system capabilities are sufficient for this short time, therefore migration is often less costly when compared to coexistence.

Migration Prerequisites

Before implementing any collaboration migration steps, all administrators should ensure their IP infrastructure is "collaboration ready," including redundancy, high availability, power consumption, Quality of Service (QoS), in-line power, ethernet ports, and so forth. For further details, refer to the chapter on Network Infrastructure, page 3-1.

If this is a first deployment of Cisco Unified Communications on Cisco Unified Computing System (UCS), follow the guidance provided in the *Cisco UCS Site Preparation Guide*, available at

https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/hw/site-prep-guide/ucs_site_prep.html

Business needs of the users play an important role in identifying key system requirements to ensure that the features and functionality are reserved or translated during migration to provide equivalent behavior. A list of features and the versions of various devices and software helps in understanding what is supported. Typically some kind of survey of the site and users should be performed to ensure that all requirements (for example, fax/modems, environmental control systems, and so forth) are appropriately identified and accounted for.

Cisco Collaboration System Migration

There are two main methods for migrating to a virtualized Cisco Collaboration System: phased migration and parallel cutover. Cisco Prime Collaboration Deployment can be used to manage and simplify the migration process.

Phased Migration

This method typically starts with a small trial focused around Cisco Unified Communications Manager (Unified CM). Once the customer is familiar with Unified CM, the system administrator can initiate the steps to migrate and move groups of users at a time to the production system, with the new Unified CM release.

Parallel Cutover

This method begins similar to the phased approach; however, once the customer is satisfied with the progress of the trial, then a time and date are chosen for cutting-over all the users at once to the new Cisco Collaboration System.

A parallel cutover has the following advantages over a phased migration:

- If something unexpected occurs, the parallel cutover provides a back-out plan that allows you to revert, with minimal effort, to the previous system, which is essentially still intact. For example, with phased migration from a PBX, service can be restored to the users simply by transferring the inbound PSTN trunks from the Cisco Collaboration System gateway(s) back to the PBX.
- Parallel cutover allows for verification of the configuration of the collaboration services before the system carries live traffic. This scenario can be run for any length of time prior to the cutover of the collaboration services, thereby ensuring correct configuration of all user information such as phones, gateways, the dial plan, mailboxes, and so forth.

- The parallel cutover allows for verification of the configuration of the Unified Communications service before the system carries live traffic. This scenario can be run for any length of time prior to the cutover of the Unified Communications service, thereby ensuring correct configuration of all user information such as phones, gateways, the dial plan, mailboxes, and so forth.
- Training can be carried out at a more relaxed pace by allowing subscribers to explore and use the collaboration services at their own pace prior to the cutover.
- The system administrator does not have to make special provisions for "communities of interest." With a phased approach, you have to consider maintaining the integrity of features such as call pickup groups, hunt groups, shared lines, and so forth. These associations can be easily accounted for when moving the complete service in a parallel cutover.

One disadvantage of the parallel cutover is that it requires the collaboration services, including the supporting infrastructure, to be fully funded from the beginning because the entire system must be deployed prior to bringing it into service. With a phased migration, on the other hand, you can purchase individual components of the system as and when they are needed, and this approach does not prevent you from starting with a small trial system prior to moving to full deployment. Neither method is right or wrong, and both depend upon individual customer circumstances and preferences to determine which option is most suitable.

Cisco Collaboration System Migration Examples

The following examples illustrate a phased migration and a parallel cutover from a PBX system to a Cisco Collaboration System.

Example 26-1 Phased Migration to Cisco Collaboration System

This approach typically entails a small Cisco Collaboration System trial that is connected to the main corporate PBX. The choice of which signaling protocol to use is determined by the required features and functionality as well as by the cost of implementation. Cisco Unified Communications Manager (Unified CM) can support either regular PSTN-type PRI or QSIG PRI as well as H.323 and SIP. Of these options, QSIG PRI typically provides the highest level of feature transparency between any two systems.

PSTN-type PRI provides for basic call connectivity as well as Automatic Number Identification (ANI). In some instances, the protocol also supports calling name information. This level of connectivity is available to all PBXs and therefore is considered to be the least costly option; that is, if the PBX can connect to the public network through PRI, then it can connect to Unified CM because Unified CM can be configured as the "network" side of the connection.

With either PSTN-type PRI or QSIG, the process for a phased migration is similar: move users from the PBX to Unified CM in groups, one group at a time, until the migration is complete.

The Cisco San Jose campus, consisting of some 23,000 users housed in approximately 60 buildings, was migrated to a Cisco Collaboration System in this manner and took just over one year from start to finish at the rate of one building per weekend. All users in the selected building were identified, and their extensions were deleted from the PBX on a Friday evening. At the same time, additions were made to the PBX routing tables so that anyone dialing those extension numbers would then be routed over the correct PRI trunk for delivery to Unified CM. During the weekend, new extensions were created in Unified CM for the users, and new IP phones were delivered to their appropriate office locations, ready for use by Monday morning. This process was repeated for each building until all users had been migrated.

Example 26-2 Parallel Cutover to Cisco Collaboration System

For this approach, all IP phones and gateways are fully configured and deployed so that users have two phones on their desk simultaneously, an IP phone as well as a PBX phone. This approach provides the opportunity not only to test the system but also to familiarize users with their new IP phones. Outbound-only trunks can also be connected to the Cisco Collaboration System, giving users the opportunity to use their new IP phones to place external as well as internal calls.

Once the Cisco Collaboration System is fully deployed, you can select a time and date for bringing the new system into full service by transferring the inbound PSTN trunks from the PBX to the Cisco Collaboration System gateways. You can also leave the PBX in place until such time as you are confident in the operation of the Cisco Collaboration System, at which point the PBX can then be decommissioned.

The Cisco San Jose campus voicemail service was originally provided by four Octel 350 systems serving some 23,000 users. Cisco Unity servers were installed and users' mailboxes were configured. Users had access to the their Cisco Unity mailbox by dialing the new access number, in order to allow them to record their name and greeting(s) as well as to familiarize themselves with the new Telephony User Interface (TUI). Approximately two weeks later, a Unified CM Bulk Administration Tool (BAT) update was carried out on a Friday evening to change the Call-Forward Busy and No-Answer (CFB/CFNA) numbers as well as the Messages button destination number for all users to the Unity system. Upon returning to work on Monday morning, users were serviced by Cisco Unity. The Octel 350 systems were left in place for one month to allow users to respond to any messages residing on those systems before they were decommissioned.

Summary of Cisco Collaboration System Migration

Although both methods of Cisco Collaboration System migration work well and neither method is right or wrong, the parallel cutover method usually works best in most cases. Cisco has a lab facility dedicated to testing interoperability between Unified CM and PBX systems, which might or might not include your current system architecture and applications. Cisco documents these test systems and their interoperability and end-user features, which should help greatly with the installation and migration process. The results of that testing are made available as application notes, which are posted at

https://www.cisco.com/c/en/us/solutions/enterprise/interoperability-portal/index.html

The notes are updated frequently, and new documents are continuously added to this website. Check the website often to obtain the latest information.

Cisco Prime Collaboration Deployment is the primary tool for migration from a Cisco Collaboration System running on MCS servers to a Cisco system running with virtualization, or migration from a previous system version to Cisco Collaboration System Release 12.*x*.

Centralized Deployment

In the case of an enterprise that has chosen a centralized deployment model for its Cisco Collaboration System, two options exist:

- Start from the outside and work inward toward the central site (that is, smallest to largest).
- Start from the central site and work outward toward the edges.

The majority of customers choose the first option because it has the following advantages:

- It provides the opportunity to fully deploy all the Cisco Collaboration services and then conduct a small trial prior to rolling the services out to the remote locations.
- The rollout of Cisco Collaboration services can be done one location at a time, and subsequent locations can be migrated whenever convenient.
- This option is the lowest cost to implement once the core Cisco Collaboration services are deployed at the central site.
- IT staff will gain valuable experience during migration of the smaller sites prior to migrating the central site.

The remote sites should be migrated by the parallel cutover approach, whereas the central site can be migrated using either the parallel or phased approach.

Which Cisco Collaboration Service to Migrate First

This choice depends mainly on the customer's particular business needs, and Cisco Collaboration solutions allows for most of the individual services to be deployed independently of the others. For example, IP telephony and voice messaging can be deployed independently from each others.

This capability provides the customer with great flexibility. Consider a customer who is faced with a voicemail system that has since gone end-of-support and is suffering various issues leading to customer dissatisfaction. Cisco Unity Connection can often be deployed and integrated with the current PBX, thereby solving this issue. Once the new voicemail system is operating appropriately, then attention can turn to the next collaboration service, namely IP telephony.

Migrating Video Devices to Unified CM

Video endpoints controlled by Cisco Unified Communications Manager (Unified CM) might support only a subset of the features that are available with a video-centric Cisco Video Communications Server (VCS). However, migration to Cisco Unified CM can provide advantages such as a unified dial plan and consolidation of other features under a single call control agent. The following guidelines apply for migrating video endpoints to Unified CM:

- Ensure that the technical functionality (for example, codecs or the ability to do content sharing) is fully supported so that the migration will not result in the loss of any features.
  - Phased migration is the most commonly used method for this purpose because it enables users to become familiar with the new devices while still having their existing phones as backup for some time.
  - Provide adequate network capacity to ensure a good experience for users. As the video resolution increases, higher bandwidth is needed when compared with audio-only calls.
  - Migrate the dial plan and associated gateways (for example, ISDN H.320 gateways) and application servers (for example, conferencing servers and bridges).
- For endpoints, consider any additional licenses needed if endpoint versions will be upgraded or if some devices need different licenses.
  - System management tools can be a big help when there is a large number of endpoints or if the endpoints need more back-end administration and support.

Organizations can then assess the types of devices, the feasibility, and the scope of the tasks needed so that the migration of video devices to Unified CM is as efficient and effective as possible.

Migrating Licenses to Cisco Collaboration System Release 12.*x*

Cisco Collaboration System Release 12.*x* provides centralized license management through Cisco Smart Software licensing. The licensing model for 12.*x* releases is under the management of the Cisco Smart Software Manager (SSM) instead of Prime License Manager as with previous system releases.

Customers who have already deployed Cisco Unified Communications or Collaboration solutions can use the Cisco Global Licensing Operations (GLO) process to migrate existing licenses to Cisco Smart Software licensing.

License Migration with Cisco Global Licensing Operations (GLO)

A self-serve option is available for those experienced with the Cisco software licensing process and functions. This self-serve option is available through the Product License Registration tool at

https://tools.cisco.com/SWIFT/LicensingUI/migrateDisplayProducts

Note You must have a valid Cisco.com login ID and password in order to access the Product License Registration tool at the above link.

You may also get assistance from the Cisco Global Licensing Operations (GLO) team by opening a case using the Support Case Manager https://tools.cisco.com/ServiceRequestTool/scm/mgmt/case.

For all license migrations, follow the guidelines presented in this section.

The license migration process has been simplified to make the migration to Cisco Smart Software License Manager much easier. Customers wanting to upgrade to Cisco Collaboration System Release 12.*x* may contact the Cisco Global Licensing Operations (GLO) team directly for all migration needs. GLO processes the request and transfers the licenses to the organization's Cisco Smart Account. Cisco adjusts the current software service contract product records to reflect the number of Release 12.*x* users that have been licensed.

Make sure you register any unused Product Activation Keys (PAKs) for the system being migrated. If the customer had planned for growth in the previous license model, take that into consideration for the current migration. As an example, if there is a need to have some clusters on a pre-12.*x* release while upgrading the rest to 12.*x*, then the license migration request to the GLO team must clearly state exactly how many users need to be migrated and how many will remain in the old format when the request is made.

Ensure that you have analyzed all licensing needs in detail, and clearly state the needs in the request. If migration is from a pre-9.*x* release with DLUs, once those DLUs are migrated there is no way to revert back to the old schema because all migrated DLUs will be in the "revoked" state.

Note The licensing process is subject to change with each new release. Always confirm the process with Cisco GLO before submitting your license request to license@cisco.com.

You can contact GLO for license migration during the following stages:

Before the Upgrade

You will need to provide GLO with the following information:

- If you are upgrading from a pre-9.*x* system, then use License Count Utility (LCU) output that was run on the Cisco Unified Communications Manager publisher node.
- MAC address of the Cisco Unified Communications Manager publisher node. If available, include all previous publisher or license MAC addresses.

After the Upgrade

You will need to provide GLO with the following information:

- If you are upgrading from a pre-9.*x* system, then use License Count Utility (LCU) output that was run on the Cisco Unified Communications Manager publisher node.
- MAC address of the Cisco Unified Communications Manager publisher node. If available, include all previous publisher or license MAC addresses.
- Site information for the contract update (for example: name-all name permutations, city, state, country)
- (Optional) Email addresses to send license and software support contract updates.
- (Optional) If a User Connect License (UCL) customer, how the customer wants to allocate unused DLUs.

Note For all pre-9.*x* systems moving to Collaboration System Release 12.*x*, customers must decide if they want to use their unused DLUs or drop them at the time of migration. There are no refunds for dropped DLUs; however, customers will save on future service charges. Note the differences between current Cisco Unified Communications Software Subscription (UCSS) users on contract and estimate the change, if any, in their UCSS and Essential Operate Services (ESW) costs at renewal.

At the time of migration, customers may choose how to use their existing licenses. The options are:

- Keep the same quantity and type of licenses.
- Decrease license quantity and change type (no refund).
- Increase license quantity by converting DLUs.

After the upgrade process is complete, the information is locked in and becomes the customer entitlement record moving forward. There are no further modifications to the license migration information.

For more information, refer to the following documentation:

- *Cisco Smart Software Licensing Overview*, available at

  https://www.cisco.com/go/smartlicensing

- Cisco Smart Software Licensing information in the latest version of the *System Configuration Guide for Cisco Unified Communications Manager*, available at

  https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Cisco Smart Software Manager

Cisco Smart Software Manager (Cisco SSM) currently supports the following Cisco Collaboration applications:

- Cisco Unified CM
- Cisco IM and Presence Service
- Cisco Unity Connection
- Cisco Emergency Responder

For more information on Cisco Smart Software Manager, refer to the documentation available at

https://www.cisco.com/c/en/us/buy/smart-accounts/software-manager.html

Using Cisco Prime Collaboration Deployment for Migration from Physical Servers to Virtual Machines

Cisco Prime Collaboration Deployment is a management application that enables administrators to perform migration from legacy Cisco Unified CM and Cisco IM and Presence services to a virtualized environment in Cisco Collaboration System Release 12.*x*. Cisco Prime Collaboration Deployment can migrate the cluster(s), handle data migration, and install the 12.*x* release on all the new VMware ESXi hosts with very little impact to the production (source) cluster. Cisco Prime Collaboration Deployment does a direct migration, whereas previous migration methods involved more steps with a "server recovery" Disaster Recovery System relying on an initial upgrade followed by a restore from backup.

Cisco Prime Collaboration Deployment can also be used to:

- Upgrade Unified Communications software (8.6.1 and later releases)
- Install Cisco Option Package (COP) files (locales or device packs) on a cluster (8.6.1 and later releases)
- Switch versions
- Reboot
- Change IP addresses or hostnames on existing clusters
- Install a new Unified CM or IM and Presence cluster

Cisco Prime Collaboration Deployment Migration Types

Cisco Prime Collaboration Deployment supports two types of migrations:

- Simple Migration, page 26-10

  In a simple migration, each node in the cluster keeps its original hostname and IP address as well as all other network configurations.

- Network Migration, page 26-10

  In a network migration, one or more nodes in the cluster have a change in hostname, IP address, and/or any other network configuration needed for Collaboration Applications.

Cisco Prime Collaboration Deployment Migration Prerequisites

- Install the VMware ESXi Hypervisor.
- Deploy the Cisco Prime Collaboration Deployment virtual machine (delivered as a virtual appliance).
- Download the Open Virtualization Archive (OVA) files for the Cisco Collaboration applications, and create target virtual machines using the OVA.
- Download Cisco ISO images for the target release and upload them to Cisco Prime Collaboration Deployment.
- Install Cisco Collaboration System Release 12.*x* nodes on the virtual machines.

Simple Migration

With this type of migration, the IP addresses and hostnames are not changed during the migration. The following procedure describes the recommended steps for migration task configuration within Cisco Prime Collaboration Deployment to ensure a highly available migration.

Cisco Prime Collaboration Deployment first exports the data of all the existing nodes. Then it powers down the existing publisher, installs the new publisher running as a virtual machine, and imports the publisher's data.

After the publisher migration is done, Cisco Prime Collaboration Deployment migrates the TFTP and backup call processing nodes of the cluster. First the existing TFTP and backup call processing nodes are powered down. Then Prime Collaboration Deployment installs the new TFTP and backup call processing nodes, and imports the backup data.

Once the backup call processing nodes are migrated, the Cisco Prime Collaboration Deployment migration task pauses and the administrator should configure all phones to re-register to the backup call processing nodes by changing the order in the Unified Communications Manager group, or by using device pools.

Finally, Cisco Prime Collaboration Deployment migrates the primary call processing nodes. Once this is done, phones can be re-registered to their primary call processing server.

For more details, refer to the latest version of the *Upgrade and Migration Guide for Cisco Unified Communications Manager and IM and Presence Service*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-guides-list.html

Network Migration

Cisco Prime Collaboration Deployment can also be utilized for a migration requiring network configuration changes for parameters such as server IP address and hostname. If the source Unified CM cluster is running release 8.*x* or later, the phone Initial Trust List (ITL) file should be updated during the migration by using the Bulk Certificate Management export, consolidate, and import functions in order to avoid having to delete the ITL file manually on each phone.

For more information, refer to the latest version of the *Cisco Prime Collaboration Deployment Administration Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

Migrating Video Endpoints from Cisco VCS to Unified CM

Video endpoints migrated from the Cisco TelePresence Video Communication Server (VCS) to Cisco Unified CM might support only a subset of the features that were available in the video-centric VCS environment. However, migration to Cisco Unified CM can provide advantages such as a unified dial plan and consolidation of other features under a single call control agent. Migration of video endpoints to Cisco Unified CM supports both SIP and SCCP endpoints, but Cisco recommends that all customers move to SIP endpoints. Although SCCP is supported, SIP has grown in popularity among both Unified Communications vendors and customers, and SIP features and functionality have grown, making SIP the new standard and recommended choice for Unified Communications.

Consider the following recommendations when migrating video endpoints to Unified CM as SIP endpoints:

- Ensure that the technical functionality (for example, codecs or the ability to do content sharing) is fully supported so that the migration will not result in the loss of any features.
- Provide adequate network capacity to ensure a good experience for users. As the video resolution increases, higher bandwidth is needed when compared with audio-only calls.
- Migrate the dial plan and associated gateways or trunks (for example, ISDN H.320 gateways) and application servers (for example, conferencing servers and bridges).
- For endpoints, consider any additional licenses needed if endpoint versions will be upgraded or if some devices need different licenses.
- System management tools can be a big help when there is a large number of endpoints or if the endpoints need more back-end administration and support.
- Customers should assess the types of devices, the feasibility, and the scope of the tasks needed so that the migration of video devices to Unified CM is as efficient and effective as possible.

Migrating from H.323 to SIP

H.323 was designed with a good understanding of the requirements for multimedia communications over IP networks, including audio, video, and data conferencing. It defines an entire unified system for performing these functions. Although SIP adoption is on the rise, H.323 is still the most widely deployed protocol for videoconferencing endpoints due to its longevity in the field. Organizations have spent a lot of effort and money deploying H.323, so they understand how it fits into their environment.

SIP is easier to implement and has begun to gain popularity in the video marketplace. As organizations struggle with the idea of changing signaling protocols, the industry has continued to evolve, and SIP has become popular for its ease of use and ability to integrate with other vendors' products. Cisco Collaboration Systems support both H.323 and SIP, but Cisco strongly recommends utilizing SIP because it provides a set of services similar to H.323 with far less complexity, rich extensibility, and better scalability.

Migrating Trunks from H.323 to SIP

Cisco Unified CM supports both SIP and H.323 intercluster trunks, and in many cases the decision to use SIP or H.323 is driven by the unique feature(s) offered by each of the protocols. A number of factors can dictate the choice customers make, such as experience, ease of interoperability, features, and

functionality with various other products. While Cisco Collaboration Systems support both H.323 and SIP trunks, Cisco recommends using SIP trunks for all deployments because SIP trunks provide ease of interoperability as well as additional features and functionality not available with H.323 trunks.

For detailed information on H.323 and SIP trunk capabilities and operation, see the chapter on Cisco Unified CM Trunks, page 6-1.

Migrating Gateways from H.323 to SIP

Cisco Unified Communications Manager (Unified CM) supports both SIP and H.323 protocols for gateways. Cisco gateways provide a number of methods for connecting Cisco Collaboration Systems to the Public Switched Telephone Network (PSTN), a legacy PBX, or third-party external deployment solution. Cisco provides both voice and video gateways, from entry-level to high-end, that fully support both protocols, but Cisco highly recommends SIP as the choice for all call signaling because it interoperates better with the entire portfolio of Cisco Collaboration voice and video products.

For detailed information on H.323 and SIP gateway capabilities and operation, see the chapter on Gateways, page 5-1.

Migrating Endpoints from SCCP to SIP

Given Cisco's general recommendation for Session Initiation Protocol (SIP) standard signaling, administrators should consider migrating their existing SCCP IP endpoints to SIP IP endpoints to provide feature parity and standards compliance. If existing SCCP IP phone models support SIP loads, administrators can use the Cisco Unified CM Bulk Administration Tool (BAT) to perform this migration.

Unified CM BAT is the recommended tool to migrate SCCP phones to SIP phones. SIP is a universal protocol standard in the industry. Within Unified CM BAT, there is an option for SCCP-to-SIP phone migration workflow (**Bulk Administration > Phones > Migrate Phones > SCCP to SIP**), which generates a report of existing SCCP phones. After selecting the SCCP phones to be migrated, the job to migrate these devices to SIP can be run immediately or scheduled for later. During the SCCP-to-SIP migration, only SIP specific default values in the phone report get migrated; other values in the template are not migrated. Migrating a phone from SCCP to SIP does not require a manual reset because the migration itself handles the reset of the phones.

For more information and detailed migration steps, refer to the latest version of the *Bulk Administration Guide for Cisco Unified Communications Manager*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

SIP URI Dialing and Directory Numbers

SIP uniform resource identifiers (URIs) are an addressing schema for directing calls to users. A URI is essentially an alias for a user's assigned directory number. A SIP URI resembles an email address and is written in the following format:

sip:*x*@*y*:*Port*

where *x*=username and *y*=host (domain or IP)

For example:

username@cisco.com or users-directorynumber@cisco.com

URIs are alphanumeric strings consisting of a user name and a host address separated by the @ symbol. Cisco Unified CM supports the following formats for directory URIs:

- *User@domain* (for example, joe@cisco.com)
- *Users-directorynumber@domain* (for example, 9728135555@cisco.com)

If the SIP request carries a user=phone tag, the SIP URI is interpreted as a numeric SIP URI and Unified CM assumes that the user portion of the SIP URI is a directory number. If no user=phone is present, the decision is based on the dial string interpretation setting in the calling device's (endpoint or trunk) SIP profile. This setting either defines a set of characters that Unified CM will accept as part of numeric SIP URIs (0-9, *, #, +, and optionally A-D) or it enforces the interpretation as a directory URI.

Note If you do not specify a port, the default SIP port (5060) is assumed. If you have changed the default SIP port to something else, then you must specify it in the SIP URI.

The following URI and directory number (DN) considerations apply to Unified CM and supported endpoints:

- DNs are the primary identity of the endpoint device.
- URIs are assigned to DNs, and each DN can support up to 5 URIs.
- Devices always register with their DNs.
- URIs can be dialed from Cisco Unified IP Phone 9900 Series, Cisco Unified IP Phone 8961, Cisco Jabber, Cisco DX Series, and third-party endpoints.
- The primary URI can be synchronized directly from LDAP in Unified CM.

For more information, see the section on Implementing Endpoint SIP URIs, page 14-78.

USB Support with Virtualized Unified CM

Music on hold (MoH) from an audio source file (using unicast or multicast) is supported; however, fixed or live audio source connections to Unified CM are not supported due to lack of USB audio support with virtualized servers. The following guidelines apply to live audio source MoH with virtualized Unified CM server nodes:

- Live or fixed audio source feeds from a USB audio device are not supported on Unified CM.
- Instead, a Cisco IOS router may be used to deliver multicast MoH feeds from fixed or live audio sources. This requires configuration of multicast MoH on the Cisco IOS router using Survivable Remote Site Telephony (SRST) or Enhanced SRST.
- Multicast must be enabled on the network to enable the Cisco IOS router to stream audio to endpoints and gateways.

While the Cisco UCS B-Series Blade Servers and C-Series Rack-Mount Servers do support a local keyboard, video, and mouse (KVM) cable connection that provides a serial port, a Video Graphics Array (VGA) monitor port, and two Universal Serial Bus (USB) ports, the Unified CM VMware virtual application has no access to these USB and serial ports. Therefore, Unified CM no longer supports the Cisco Messaging Interface (CMI) service for Simplified Message Desk Interface (SMDI) integrations, fixed MoH audio source integration for live MoH audio feeds using the audio cards, or flash drives to these servers.

On-Premises Cisco IM and Presence Service Migration

For existing Cisco IM and Presence deployments running on a Cisco MCS physical server, upgrading to release 12.*x* would require a migration to virtualization. Use Cisco Prime Collaboration Deployment to facilitate this migration.

Network Management

Revised: March 1, 2018

Network management is a service consisting of a wide variety of tools, applications, and products to assist network system administrators in provisioning, operating, monitoring and maintaining new and existing network deployments. A network administrator faces many challenges when deploying and configuring network devices and when operating, monitoring, and reporting on the health of the network infrastructure and components such as routers, servers, switches and so forth. Network management helps system administrators monitor each network device and network activity so that they can isolate and investigate problems in a timely manner for better performance and productivity.

With the convergence of rich media and data, the need for unified management is greater than ever. The Cisco Prime Collaboration (Prime Collaboration) offers a set of integrated tools that help to test, deploy, and monitor Cisco Unified Communications and TelePresence systems. Prime Collaboration implements the various management phases to strategically manage the performance and availability of Cisco Unified Communications applications including voice, video, contact center, and rich media applications. The network management phases typically include: plan, design, implement, and operate (PDIO). Table 27-1 lists the PDIO phases and the major tasks involved with each phase.

Table 27-1 Network Management Phases and Tasks

Plan & Design	Implement	Operate
Assess the network infrastructure for Cisco Unified Communications capability. For example, predict overall call quality. Prepare the network to support Cisco Unified Communications. Analyze network management best practices.	Deploy and provision Cisco Unified Communications. For example, configure the dial plan, partitioning, user features, and so forth. Enable features and functionality on the existing infrastructure to support Cisco Unified Communications.	Manage changes for users, services, IP phones, and so forth. Generate reports for operations, capacity planning, executive summaries, and so forth. Track and report on user experiences. For example, use sensors to monitor voice quality. Monitor and diagnose problems such as network failures, device failures, call routing issues, and so forth.

This chapter provides the design guidance for the following management tools and products that fit into the implementation and operation phases of Cisco Unified Communications Management:

- Cisco Prime Collaboration manages provisioning of initial deployments and ongoing operational activation for Unified Communications and TelePresence services. Cisco Prime Collaboration provides comprehensive monitoring with proactive and reactive diagnostics for the entire Cisco Unified Communications system. It also provides a reliable method of monitoring and evaluating voice quality in Cisco Unified Communications systems. For details, refer to the related product documentation available at

  https://www.cisco.com/c/en/us/products/cloud-systems-management/prime-collaboration/index.html

- Cisco TelePresence Management Suite (TMS) offers visibility and centralized control of your telepresence videoconferencing network, including remote systems. For details, refer to the related product documentation available at

  https://www.cisco.com/c/en/us/products/conferencing/telepresence-management-suite-tms/index.html

For information on which software versions are supported with Cisco Unified Communications Manager (Unified CM), refer to the latest version of the *Compatibility Matrix for Cisco Unified Communications Manager and the IM and Presence Service*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-device-support-tables-list.html

What's New in This Chapter

Table 27-2 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 27-2 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Cisco Prime License Manager has been replaced by Cisco Smart Software Licensing	Cisco Smart Software Licensing, page 27-21	March 1, 2018
Other minor corrections and updates	Various sections of this chapter	March 1, 2018

Cisco Prime Collaboration

Cisco Prime Collaboration provides comprehensive voice and video network monitoring with diagnostics for the Cisco Collaboration systems, including the underlying transport infrastructure. Prime Collaboration is a converged application that eliminates the need to manage the video deployments separately from voice. It is delivered as two separate applications, Prime Collaboration Assurance and Prime Collaboration Provisioning, which are installed on separate virtual machines. Prime Collaboration is available in two modes: Standard and Advanced mode.

Prime Collaboration Assurance application provides:

- End-to-end service monitoring for Cisco Collaboration applications
- Real-time service troubleshooting and diagnostics for Cisco TelePresence systems and phones
- Video service readiness assessment

- Diagnostics tests using Cisco IP Service Level Agreements (IP SLA) and Video SLA Assessment Agent (VSAA)
- Service-level and inventory reports for voice and video systems

Note Prime Collaboration Assurance Advanced also includes Prime Collaboration Analytics. If you have purchased the Prime Collaboration Analytics license, you can access the Prime Collaboration Analytics dashboards. Prime Collaboration Analytics helps you identify traffic trends, technology adoption trends, and over/under-utilized resources in your network. You can also track intermittent and recurring network issues and address service quality issues.

Prime Collaboration Provisioning application provides:

- Standard services (phone, line, and voicemail, for example) to be ordered for subscribers (the owner of the individual phone, voicemail, or other service)
- Configuration templates that provide the ability to auto-configure the Cisco Unified Communications voice infrastructure in a consistent way
- Easy addition of the Provisioning application to an existing Cisco Unified Communications network
- Simplified policy-driven Day 2 provisioning interface to move, add, delete, or change phone users
- A Self-Care feature that enables end users to modify personal options quickly and easily

For information on the benefits and key features of Prime Collaboration and guidelines for deployment (white papers), refer to the Cisco Prime Collaboration documentation available at

https://www.cisco.com/go/primecollaboration

Failover and Redundancy

Prime Collaboration does not currently support failover. However, it can support Network Fault Tolerance when deployed on server platforms with dual Ethernet network interface cards (NICs) that support NIC teaming. This feature allows a server to be connected to the Ethernet through two NICs and, therefore, two cables. NIC teaming prevents network downtime by transferring the workload from the failed port to the working port. NIC teaming cannot be used for load balancing or for increasing the interface speed.

Prime Collaboration Assurance provides geographic redundancy through the use of VMware vSphere replication. It requires VMware activation at remote sites only.

Cisco Prime Collaboration Server Performance

Prime Collaboration runs only in a virtual environment and it requires a minimum of one virtual machine per component. If you want Assurance and Provisioning, you will need two virtual machines (one for each). For specific system requirements and capacity information, refer to the latest version of the *Cisco Prime Collaboration Provisioning Install and Upgrade Guide*, available at

https://www.cisco.com/c/en/us/support/cloud-systems-management/prime-collaboration/products-installation-guides-list.html

Network Infrastructure Requirements for Cisco Collaboration and Network Management Applications

Cisco highly recommends that you enable Domain Name Service (DNS) in the network to perform a reverse lookup on the IP address of the device to get the hostname for the device. If DNS is not desired, then host files may be used for IP address-to-hostname resolution.

Network Time Protocol (NTP) must be implemented to allow network devices to synchronize their clocks to a network time server or network-capable clock. NTP is a critical network service for network operation and management because it ensures accurate time-stamps within all logs, traps, polling, and reports on devices throughout the network.

You should enable Cisco Discovery Protocol (CDP) within the network to ensure proper monitoring. Prime Collaboration's automated device discovery is based on a CDP table. Ping Sweep may be used instead of CDP, but IP phones discovered using Ping Sweep are reported in "unmanaged" state. Simple Network Management Protocol (SNMP) must also be enabled on network devices to allow Prime Collaboration to get information on network devices at configured polling intervals and to receive alerts and faults via trap notification sent by the managed devices.

For more information on Cisco Unified Communications network requirements, see the chapter on Network Infrastructure, page 3-1.

Assurance

Cisco Prime Collaboration Assurance is a comprehensive video and voice service assurance and management system with a set of monitoring, troubleshooting, and reporting capabilities that help ensure end users receive a consistent, high-quality video and voice collaboration experience. Prime Collaboration Assurance is available in two modes: Standard and Advanced.

Prime Collaboration Advanced provides all the features that enable integrated assurance management of applications and the underlying transport infrastructure. This includes real-time monitoring and troubleshooting of Cisco TelePresence solutions and the entire Unified Communications system.

Prime Collaboration Standard provides basic assurance features that help you manage Unified Communications and TelePresence components. The features include:

- Support for Unified Communications components including voicemail and IM and Presence
- Fault monitoring for core Unified Communications components (Cisco Unified CM and Cisco Unity Connection)
- Pre-configured and customizable performance metrics dashboards that display term trends for core Unified Communications components
- Support for TelePresence components, including Cisco TelePresence Video Communication Server (VCS)
- Contextual cross-launch of serviceability pages of Unified Communications components
- Single-level role-based access control (RBAC)

Prime Collaboration Standard also includes the following features to help you manage the Unified Communications and TelePresence components:

- Device Inventory Management

  You can discover and manage endpoints that are registered to Cisco Unified Communications Manager (phones and TelePresence endpoints), Cisco TelePresence VCS, and Cisco TelePresence Management Suite (TMS). As part of the discovery, the device details are also retrieved and stored in the Prime Collaboration database. After the discovery is complete, you can perform the following device management tasks:

  - Add or remove devices
  - Manage device credentials
  - Discover devices

- Monitoring and Fault Management

  Service operators need to quickly isolate the source of any service degradation in the network for all voice and video sessions in an enterprise. Prime Collaboration provides a detailed analysis of the service infrastructure and network-related issues.

  Prime Collaboration periodically imports information from the managed devices based on the polling parameters you configure.

  The Home page includes several pre-configured dashlets that help you monitor system performance, device status, device discovery, CTI applications, and voice messaging ports. These dashlets enable you to monitor a set of predefined management objects that monitor the health of the system. From the dashlets, you can launch contextual serviceability pages.

  Prime Collaboration ensures near real-time quick and accurate fault detection. Prime Collaboration enables you to monitor the events that are of importance to you. You can set up Prime Collaboration to send notifications for alarms.

  In addition to the faults that are present in the Cisco TelePresence Management System and Unified Communications applications, it also displays the custom tickets that are raised on Cisco TMS.

  Using the Alarm browser, you can view the alarms and events in the system and initiate troubleshooting. You can also configure Prime Collaboration to send fault notifications, and you can view call connection/disconnection details related to the Cisco TMS applications in the Call Events UI.

Cisco Prime Collaboration Assurance provides a unified view of the entire Cisco Unified Communications infrastructure and presents the current operational status of each element of the Cisco Unified Communications network. Prime Collaboration also provides diagnostic capabilities for faster problem isolation and resolution. In addition to monitoring Cisco gateways, routers, and switches, Prime Collaboration continuously monitors the operational status of various Cisco Unified Communications elements such as:

- Cisco Unified Communications Manager (Unified CM)
- Cisco Unified Communications Manager Express (Unified CME)
- Cisco Unified Communications Manager Session Management Edition
- Cisco Unity Connection
- Cisco Unity Express

- Cisco Unified Contact Center Enterprise (Unified CCE), Unified Contact Center Express (Unified CCX), and Unified Customer Voice Portal (Unified CVP)

Note Cisco Prime Collaboration Service Level View does not support multiple Cisco Unified Contact Center Enterprise (Unified CCE) deployments.

- Cisco IM and Presence
- Cisco Emergency Responder
- Cisco Unified Border Element
- Cisco Unified Endpoints

Note Cisco Prime Collaboration supports Unified Communications and TelePresence applications running in a virtualized environment but does not provide monitoring of VMware or hardware. Use vCenter for managing VMware hosts. For Unified Computing System (UCS) B-series Blade servers, UCS Manager provides unified, embedded management of all software and hardware components in the Cisco UCS. It controls multiple chassis and manages resources for thousands of virtual machines. For UCS C-series servers, the Cisco Integrated Management Controller provides the management service.

For more information on the supported products (particularly Cisco endpoints) and versions supported by Prime Collaboration, refer to the Cisco Prime Collaboration data sheet available at

https://www.cisco.com/c/en/us/products/cloud-systems-management/prime-collaboration/index.html

One protocol that Prime Collaboration uses to monitor the Unified Communications elements is Simple Network Management Protocol (SNMP). SNMP is an application-layer protocol using UDP as the transport layer protocol. There are three key elements in SNMP managed network:

- Managed devices — Network devices that have an SNMP agent (for example, Unified CM, routers, switches, and so forth).
- Agent — A network management software module that resides in a managed device. This agent translates the local management information on the device into SNMP messages.
- Manager — Software running on a management station that contacts different agents in the network to get the management information (for example, Prime Collaboration).

The SNMP implementation supports three versions: SNMP v1, SNMP v2c, and SNMP v3. SNMP v3 supports authentication, encryption, and message integrity. SNMP v3 may be used if security is desired for management traffic. Prime Collaboration supports all three versions of SNNP. SNMP v1 and v2c read/write community strings or SNMP v3 credentials must be configured on each device for agent and manager to communicate properly. Prime Collaboration needs only SNMP read access to collect network device information.

For more information on SNMP, refer to the Cisco Prime Collaboration documentation available at

https://www.cisco.com/c/en/us/products/cloud-systems-management/prime-collaboration/index.html

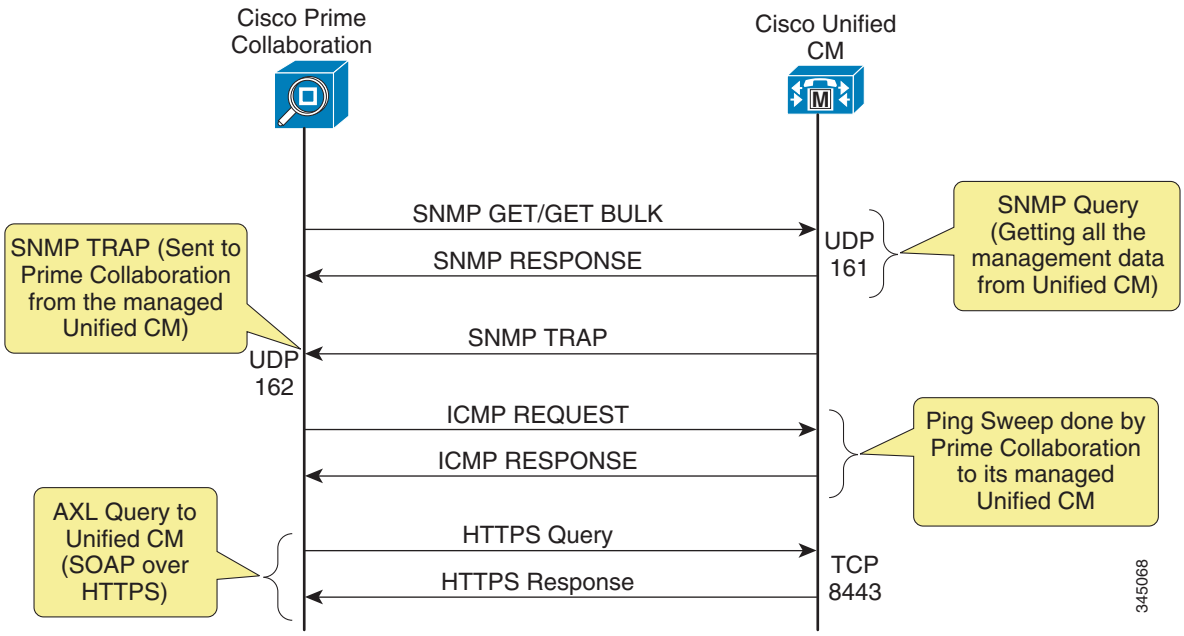
Assurance Design Considerations

Cisco Prime Collaboration interfaces with other devices in the network in the following ways:

- Simple Network Management protocol (SNMP) to manage all Cisco Unified Communications servers, gateways, and switches.
- Administrative XML Layer (AXL) to manage Unified CM. AXL is implemented as a Simple Object Access Protocol (SOAP) over HTTPS web service.
- HTTP to the IP phone to collect serial number and switch information. HTTP must be enabled on the IP phones.
- Enhanced event processing with Cisco Unified CM remote syslog integration, and leveraging the Cisco Real-Time Monitoring Tool (RTMT) interface for pre-collected Unified CM cluster-wide data
- Skinny Client Control Protocol (SCCP) and Session Initiation Protocol (SIP) to Cisco Unified IP Phones for synthetic tests.
- Internet Control Message Protocol (ICMP) or Ping Sweep for Cisco IOS routers and switches, and for other voice as well as non-voice devices.

Figure 27-1 shows the system-level overview of how Prime Collaboration leverages multiple interfaces with Unified CM to gather performance counters and alarms.

Figure 27-1 Prime Collaboration and Unified CM System-Level Integration



Call Quality Monitoring (Service Experience)

Cisco Prime Collaboration Assurance Advanced monitors the quality of calls on the Cisco Unified Communications network. It relies on Unified CM and Network Analysis Modules (NAMs) to monitor and gather quality statistics on real calls rather than simulated calls in the network. Then it compares the collected quality statistics against predefined thresholds.

Prime Collaboration Assurance Advanced is also responsible for sending voice quality information to Cisco Prime Analytics (available only with Prime Collaboration Advanced) so that Analytics can perform call data analysis and generate reports.

Note A set of global call quality thresholds can be defined as one per supported codec type. Different thresholds can be grouped together based on the Unified CM cluster being monitored.

Voice Quality Measurement

Voice quality is the qualitative and quantitative measure of the sound and conversational quality of the IP phone call. Voice quality measurement describes and evaluates the clarity and intelligibility of voice conversations. Prime Collaboration uses the Network Analysis Module (NAM) and Unified CM to monitor and report voice quality information.

Unified CM Call Quality Monitoring

Unified CM stores end-of-call video and audio information and metrics in its call detail records (CDRs) and call management records (CMRs). The CMRs and CDRs are transferred to Prime Collaboration via Secure File Transfer Protocol (SFTP) every 60 seconds. To integrate with Unified CM, Prime Collaboration must be configured as a Billing Application Server in the Unified CM Unified Serviceability configuration web page. Up to three Billing Application Servers can be configured per Unified CM cluster. The following settings must be configured for the Billing Application Server:

- Hostname or IP address of the Prime Collaboration Assurance virtual machine
- Username and password for SFTP file transfer
- Protocol: SFTP
- Directory path on the Prime Collaboration virtual machine to which CDR and CMR files are transferred

Note Cisco Jabber and endpoints running Cisco CE or TE software do not generate end-of-call audio and video information. Thus, there are no CMRs for these endpoints.

In the past, the emphasis was on using the Cisco Voice Transmission Quality (CVTQ) algorithm as one means to monitor voice quality. CVTQ is based on the Klirrfaktor (K-factor) method to estimate the MOS value of voice calls. With Cisco CSR 12.*x*, packet counts, concealment ratios, and concealment second counters represent primary statistics because they can alert the network operator before network impairment has an audible impact or is visible through MOS. Table 27-3 describes these counters as well as metrics computed from them.

Table 27-3 Counters and Metrics to Measure Call Quality

Counter or Metric	Description
Concealment	Measures packet (frame) loss and its effect on voice quality in an impaired network
CS – concealed seconds	Number of seconds when there is some concealment (might not be audible)
SCS – severely concealed seconds	Number of seconds when loss is greater than 5% (audible)
SCSR (SCS Ratio) – SCS/Duration	Metric to measure voice quality
CSR (CS Ratio) – CS/Duration	Metric to measure network quality

As noted in Table 27-3, SCSR represents a measure of voice quality and is used by Prime Assurance to grade calls. For calls less than 20 seconds in duration, the following SCSR values are used to estimate call quality:

Grade	SCSR Value
Good	Less than 0.20
Acceptable	0.20 <= SCSR <= 0.30
Poor	Greater than 0.30

For calls of 20 seconds or longer in duration, the following SCSR values are used to estimate call quality:

Grade	SCSR Value
Good	Less than 0.03
Acceptable	0.03 <= SCSR <= 0.07
Poor	Greater than 0.07

Cisco Network Analysis Module (NAM)

Cisco NAM is a traffic analysis module that leverages Remote Monitoring (RMON) and some SNMP Management Information Bases (MIBs) to enable network administrators to view all layers of the Unified Communications infrastructure to monitor, analyze, and troubleshoot applications and network services such as QoS for voice and video applications. Voice instrumentation added in Cisco NAM 4.0 enables NAM integration with Prime Collaboration for call metrics through NAM-embedded data collection and performance analysis.

The Cisco NAM complements Prime Collaboration to deliver an enterprise-wide voice management solution. The NAM Appliances come with a graphical user interface for troubleshooting and analysis, and they provide a rich feature set for voice quality analysis with RTP and voice control and signaling monitoring.

Cisco Prime Collaboration polls the NAM every 60 seconds for voice quality metrics. It then does a MOS calculation on the data. This enables Prime Collaboration to correlate CDR and call stream reports from the NAM for enhanced analysis.

For more information on Cisco NAM, refer to the following site:

https://www.cisco.com/go/nam

Comparison of Voice Quality Monitoring Methods

Unified CM call quality (CDRs and CMRs) and NAM complement each other and provide a total solution for voice quality measurement. The following list notes key differences between voice quality monitoring with Unified CM and Cisco NAM:

- The Cisco NAM provides voice quality statistics every 60 seconds. Unified CM provides voice quality statistics after the call is completed (ended).
- Unified CM monitors only the call segment within its own cluster.
- Unified CM voice quality monitoring is best used to gauge the overall voice call quality in the network.

Even if Unified CM call quality metrics are not used, Prime Collaboration uses Unified CM CDR information to correlate with the NAM report for the following information:

- Source and/or destination extension number
- Device types
- Interface through which the call flowed in the case of a call to or from a gateway
- Call disconnect reason, where possible
- Exact Unified CM server (not just the Unified CM cluster) to which the phone is connected

Trunk Utilization

Cisco Prime Collaboration provides real-time Unified CM trunk utilization performance graphs. It is also tightly integrated with Cisco Prime Analytics in order to provide the call information it collects to Analytics for long-term trending and reporting purposes. The call information is provided from the CDR and CMR records Prime Collaboration gathers from Unified CM.

Failover and Redundancy

The Unified CM publisher server is responsible for transferring CDR and CMR files to Prime Collaboration via SFTP. If the publisher server is unavailable, there is no failover mechanism for Prime Collaboration to obtain the new CDR and CMR files that contain MOS values of calls in the Unified CM cluster.

Voice Monitoring Capabilities

Cisco Prime Collaboration supports the following voice quality monitoring capacities:

- Any of the following scenarios:
  - 5,000 sensor-based RTP streams per minute (with NAM modules)
  - 1,600 Unified CM calls per minute
  - 1,500 RTP streams and 666 Unified CM calls per minute

- Prime Collaboration automatically selects and gathers voice quality information (via CDR and CMR files) for all Cisco Unified IP Phones configured in a given Unified CM cluster. There is no configuration option to monitor only certain IP phones in the cluster.

Note When Cisco Prime Collaboration is operating at full capacity, its projected database growth (for Syslog, CDR, and CMR files) is estimated to be about 2.4 GB per day.

Assurance Ports and Protocol

Table 27-4 lists the ports used by the various protocol interfaces for Cisco Prime Collaboration for Assurance. Cisco recommends opening these ports in the corporate internal firewalls (if applicable) to allow communications between Prime Collaboration and other devices in the network

Table 27-4 Cisco Prime Collaboration Port Utilization for Assurance

Protocol	Port	Service
UDP	161	SNMP Polling
UDP	162	SNMP Traps
TCP	80	HTTP
TCP	443	HTTPS
TCP	1741	CiscoWorks HTTP server
UDP	22	SFTP
TCP	43459	Database
UDP	514	Syslog
TCP	8080	Determining status of Unified CM web service
TCP	8443	SSL port between Unified CM and Prime Collaboration

Note The Cisco NAM is accessed remotely over HTTPS with a non-default port. Prime Collaboration will authenticate with each Cisco NAM and maintain the HTTP/S session.

All the management traffic (SNMP) originating from Prime Collaboration or managed devices is marked with a default marking of DSCP 0x00 (PHB 0). The goal of network management systems is to respond to any problem or misbehavior in the network. To ensure proper and reliable monitoring, network management data must be prioritized. Implementing QoS mechanisms ensures low packet delay, low loss, and low jitter. Cisco recommends marking the network management traffic with an IP Precedence of 2, or DSCP 0x16 (PHB CS2), and providing a minimal bandwidth guarantee. The DSCP value must be configured in the Windows Operating System.

If managed devices are behind a firewall, the firewall must be configured to allow management traffic. Prime Collaboration has limited support in a network that uses Network Address Translation (NAT). It must have IP and SNMP connectivity from the Prime Collaboration server to the NAT IP addresses for the devices behind the NAT. Prime Collaboration contains static NAT support.

Bandwidth Requirements

Prime Collaboration polls the managed devices for operational status information at every configured interval, and it has the potential to contain a lot of important management data. Bandwidth must be provisioned for management data, especially if you have many managed devices over a low-speed WAN. The amount of traffic varies for different types of managed devices. For example, more management messages may be seen when monitoring Unified CM as compared to monitoring a Cisco Voice Gateway. Also, the amount of management traffic will vary if the managed devices are in a monitored or partially monitored state and if any synthetic tests are performed.

Analytics

Cisco Prime Collaboration Analytics provides many additional benefits to Prime Assurance. It provides trending to identify degradation over time. It can also utilize trending to provide capacity planning and quality of service (QoS) information. The capacity planning feature allows administrators to plan for growth and also to identify over- or under-utilized resources (for example, TelePresence endpoints) in their network. Analytics can generate automated reports that provide actionable information to CIOs and IT planners. Reports can be customized to meet unique business needs.

Analytics supports the following predefined dashboards:

- Technology Adoption
- Asset Usage
- Traffic Analysis
- Capacity Analysis
- Service Experience

Custom dashboards and dashlets can also be created if desired.

The Technology Adoption dashboard provides visibility into the progress of your voice and video deployment by showing devices deployed and minutes used. This information allows for more intelligent technology investment decisions based on current adoption analysis.

The Asset Usage dashboard shows long-term utilization trends for collaboration network resources. It provides information such as least used and most used resources such as endpoints.

The Traffic Analysis dashboard provides a means to analyze long-term service quality issues and identify voice and video traffic patterns. It offers options to show the top *N* callers, top *N* dialed numbers, top *N* off-net traffic locations, and top *N* call traffic locations.

The Capacity Analysis dashboard provides options that allow tracking of unused or under-utilized voice and video assets such as conferencing devices, call admission control bandwidth, and trunks. The information provided can assist in optimizing equipment and network costs.

The Service Experience dashboard helps identify call quality issues in the collaboration deployment. It can show top *N* endpoints with quality issues or allow filtering based on quality level. It also provides a means to analyze call failures, identify service usage by group of users or endpoints, and help effectively allocate the IT expense.

For detailed information on feature support and functionality, refer to the Cisco Prime Collaboration Analytics product documentation available at https://www.cisco.com.

Note Currently there is no redundancy or failover support with Analytics.

Analytics Server Performance

Analytics is included in the Prime Assurance OVA and runs on the same virtual machine. Note that Analytics does require a separate license.

Provisioning

Prime Collaboration Provisioning is available in the following forms:

- Prime Collaboration Provisioning Standard
- Prime Collaboration Provisioning Advanced

Prime Collaboration Provisioning Standard is a simplified version of Cisco Prime Collaboration Provisioning. It provides simplified provisioning across all collaboration services. You can provision all services including phones, voicemail clients, and video endpoints. Provisioning support is available for a single Unified Communications cluster with limited authorization roles.

Advanced Provisioning provides more advanced features such as delegation to individual domains, template support for configuring infrastructure instances, advanced batch provisioning, and so on. Table 27-5 lists the features available in Prime Collaboration Provisioning Standard and Advanced.

Table 27-5 Prime Collaboration Provisioning Standard and Advanced Features

Features	Standard	Advanced
Delegation of roles or Role-Based Access Control (RBAC)	A single user role is applicable to all domains. You cannot delegate user roles to different domains.	Any user role can be assigned to a specific logical domain based on a region or a group.
Ordering workflow roles	The ordering workflow activities (such as approving an order, assigning MAC addresses, shipping endpoints, or end user receipt of an endpoint) are not available.	The ordering workflow activities can be greatly customized based on the end user requirement. The activity roles can be enabled or disabled, and assigned to different users for an efficient ordering workflow.
Batch provisioning	Allows you to deploy a large number of services by combining them into a single batch. **Note:** Batch Provisioning is available for a single cluster only.	Provides advanced batch options such as importing users and services, and adding or modifying users and services across multiple clusters. You can also batch-import infrastructure settings across multiple clusters.
Infrastructure templates	The Infrastructure Configuration templates cannot be customized.	You can create templates to initially configure or reconfigure Cisco Unified Communications Manager, Cisco Unified Communications Manager Express, and Cisco Unity Express. You can add, edit, or delete the configuration settings, including adding or updating keywords and scheduling template provisioning.
Unified CM cluster support	You can configure a single cluster only.	You can configure multiple clusters.
API	Support for North Bound Interface (NBI) is not available.	Support for North Bound Interface (NBI) is available.

Cisco Prime Collaboration provides a simplified web-based provisioning interface for both new and existing deployments of Cisco Unified Communications Manager (Unified CM), Cisco Unified Communications Manager Express (Unified CME), Cisco Unity Connection, and Cisco Unity Express. Prime Collaboration provides provisioning for both the infrastructure and subscribers for Day 1 and Day 2 needs. Day 1 needs include configuring new deployments and adding more sites or locations; Day 2 needs include services for ongoing moves, adds, and changes on various components of the Cisco Unified Communications solution.

Cisco Prime Collaboration also exposes northbound APIs to allow Cisco and third parties to integrate with external applications such as HR systems, custom or branded user portals, other provisioning systems, and directory servers.

For details on Prime Collaboration system requirements and installation steps, provisioning users and the infrastructure of supported components, and capacity information, refer to the Cisco Prime Collaboration documentation available at

https://www.cisco.com/c/en/us/products/cloud-systems-management/prime-collaboration/index.html

To provide a better understanding of how Prime Collaboration can be used as a network management solution for provisioning various Cisco Unified Communications components, the next section presents some of the basic concepts of Prime Collaboration.

Provisioning Concepts

Cisco Prime Collaboration serves as a provisioning interface for the following components of a Cisco Unified Communications system:

- Call processors
  - Cisco Unified Communications Manager (Unified CM)
  - Cisco Unified Communications Manager Express (Unified CME)
- Message processors
  - Cisco Unity Connection
  - Cisco Unity Express
- Presence processors
  - Cisco IM and Presence
  - Cisco Voice Gateways
  - Cisco VG224, VG204, and VG202 Analog Voice Gateways

Note For more information on component version compatibility, refer to the Prime Collaboration information at https://www.cisco.com/c/en/us/products/cloud-systems-management/prime-collaboration/index.html.

The following sections describe some of the Prime Collaboration concepts involved in configuring those components.

Domain

Domains are used for administrative purposes to create multiple logical groups within a system. Domains have the following characteristics:

- A domain can be mapped to a geographical location or an organization unit.
- One domain can contain multiple call processors and multiple optional message processors.
- A given call processor or message processor can be a member of multiple domains.
- A domain can partition subscribers so that they can be administered separately.

Service Area

Service areas represent offices. Service areas determine the dial plans and other voice-related configuration settings in the domain. In reality, each office may have multiple service areas. The service area determines attributes such as device group, route partition, and calling search space used within Unified CM. Service areas have the following characteristics:

- Each service area is assigned to a single call processor and one optional message processor.
- Each service area should be associated with one dial plan.

Work Flow and Managing Orders

When deploying a new site or making moves, adds and changes to an existing site, users make all changes to the underlying systems through a two-stage process of creating an order and then processing that order. You can set policies for both of these stages. For example, you can configure the system so that one group of users can only create and submit orders, while another group of users can view and perform processing-related activities. Prime Collaboration contains an automation engine that performs the order processing, including service activation and business flow, based on how Prime Collaboration is configured.

The workflow coordinates activities of the ordering process (approval, phone assignment, shipping, and receiving).

Configuration Templates

Prime Collaboration enables you to configure Unified CM, Unified CME, Cisco Unity Express, and Cisco Unity Connection in a consistent way through the use of configuration templates. You can use these templates to configure any of these products, to perform an incremental rollout on these existing products, and to deploy a new service across existing customers.

Batch Provisioning

Creating users and provisioning their services can also be done automatically through batch provisioning for rolling out a new office or transitioning from legacy systems.

Best Practices

The following best practices and guidelines apply when using Prime Collaboration to provision Cisco Unified Communications components for any new and/or existing deployments:

- Managed devices must be up and running before using Prime Collaboration for further day-one activities such as rolling out a new site and day-two activities such as moves, adds, and changes.
- Pre-configuration is required for Cisco Unified CM, Cisco Unity Connection, Unified CME, Survivable Remote Site Telephony (SRST), Cisco Unity Express, and Cisco IM and Presence Service.

- Define the correct domains, service areas, and provisioning attributes.
- Modify only the workflow rules if necessary.
- Consider the use of Subscriber Types, Advanced Rule settings, and other configuration parameters.

The following basic tasks help support these best practices:

- Add call processors such as Unified CM, and/or Unified CME and message processors such as Cisco Unity, Unity Connection, and/or Unity Express.
- Create domains and assign call processors and message processors to the created domains.
- Provision the voice network by creating and using templates to configure Unified CMs or Unified CMEs, or import current voice infrastructure configurations from an existing deployment.
- Perform bulk synchronization of LDAP users into Prime Collaboration, if applicable.
- Set up the deployment by creating service areas for each domain (typically one per dial plan) and assigning subscriber (user) types to each service area.
- Create administrative users for each domain.
- Order, update, or change subscriber or user services.

Prime Collaboration Design Considerations

The following design considerations apply to Prime Collaboration for provisioning:

- Set up domains in one of the following ways:
  - Create a single domain for multiple sites, with multiple call processors and multiple message processors.
  - Create a domain for each site, consisting of one call processor and zero or more optional message processors.
  - Create multiple domains if different administrators are required to manage a subset of the subscribers.
- Create multiple service areas for multiple dial plans.
- Add only the Unified CM publisher as the call processor for Prime Collaboration. Any changes made to the Unified CM publisher through Prime Collaboration will be synchronized to all the Unified CM subscriber servers.
- Use configuration templates for Unified CM, Unified CME, or Cisco Unity Express.
- Use Cisco IOS commands for Unified CME and Cisco Unity Express configuration templates.
- Add Cisco Unified CM infrastructure data objects for Unified CM configuration templates.
- Change and modify the existing configuration templates for batch provisioning for large quantities of phones and lines (DNs).
- Create multiple domains if you want different domain administrators to manage different sets of subscribers for Day 2 moves, adds, and changes of services (such as phones, lines, and voicemail), even for a single-site deployment.
- Create one service area for one dial plan.
- Create multiple service areas if multiple dial plans are required for the device pools, location, calling search space, and phones.

- Prime Collaboration is an IPv6-aware application with the following characteristics:
  - Prime Collaboration communicates with Unified CM over an IPv4 link. The Prime Collaboration user configuration interface allows users to enter only IPv4 IP addresses because Unified CM has SOAP AXL interfaces in IPv4 only. Therefore, Prime Collaboration must use IPv4 addresses to communicate with the AXL interfaces on Unified CM.
  - Prime Collaboration handles the IPv6 addresses contained in SIP trunk AXL response messages.
  - Support of IPv6-aware functions does not affect support for current Cisco Unified Communications Manager Express, Cisco Unity Express, and Cisco Unity Connection devices.

Redundancy and Failover

If Prime Collaboration fails in the middle of the configuration process, changes made to the configured devices from the Prime Collaboration GUI might not be saved and cannot be restored. Administrators must use manual steps to continue the configuration process by using other tools such as telnet or login (HTTP) to the managed devices until Prime Collaboration comes back live. Manually added configuration changes to the managed device will not automatically show up in the Prime Collaboration dashboard or database unless you also perform synchronization from Prime Collaboration for the call processors (Unified CM and/or Unified CME), message processors (Unity Connection and/or Unity Express), and domains.

Provisioning Ports and Protocol

Table 27-6 lists the ports used by the various protocol interfaces for Prime Collaboration. Cisco recommends opening those ports in the corporate internal firewalls (if applicable) to allow communications between Prime Collaboration and other devices in the network.

Table 27-6 Prime Collaboration Port Utilization for Provisioning

Protocol	Port	Service
TCP	80	HTTP[1] [2]
TCP	8443	HTTPS[2]
TCP	22	SSH[3]
SSH	23	Telnet[3]
TCP	1433	Database[4]

1. To access the Prime Collaboration Administration web page.
2. Prime Collaboration provisions Unified CM via Administrative XML Layer (AXL) Simple Object Access Protocol (SOAP).
3. For Prime Collaboration to communicate with Unified CME and Cisco Unity Express.
4. For Prime Collaboration to connect to the database of Cisco Unity Connection.

Cisco TelePresence Management Suite (TMS)

Cisco TelePresence Management Suite (TMS) supports scheduling of video endpoint and conferencing devices. Scheduling ensures endpoint and port resource availability and provides convenient methods to connect to TelePresence conferences. Most organizations already use calendaring applications to schedule conferences. In this case, calendaring integration enables users to schedule conferences with their existing calendaring client.

Calendaring Options

Calendaring integration gives users the ability to schedule video conferences and invite participants directly from their calendaring application while viewing availability information of resources regardless of where meetings are created. Calendaring options include:

- Cisco TelePresence Management Suite Extension for Microsoft Exchange (TMSXE)

  Allows conference organizers to schedule conferences using their Microsoft Outlook client.

- Cisco TelePresence Management Suite Extension for IBM Lotus Notes (TMSXN)

  Allows conference organizers to schedule conferences using their IBM Lotus Notes client.

- Cisco TelePresence Management Suite Extension Booking API (TMSBA)

  Allows conference organizers to schedule conferences using additional groupware calendaring systems through API integration.

- Cisco TMS Web-based user interface

  Allows users or administrators to schedule conferences through a web-based interface. This is part of the Cisco TMS core application and does not require additional installation or integration.

For more information on Cisco TMS Extensions and APIs, refer to the product documentation available at

https://www.cisco.com/c/en/us/support/conferencing/telepresence-management-suite-extensions/tsd-products-support-series-home.html

Cisco TMS also allows user and administrators to schedule conferences through a web-based interface. This is part of the Cisco TMS core application and does not require additional installation or integrations.

Cisco highly recommends integrating your corporate calendaring application with the scheduling and management platform chosen by your organization. However, you may also choose to schedule conferences using the TMS web interfaces.

When deploying Cisco TMS calendar integration as your corporate calendaring application, choose the appropriate extension for your environment. For example, if Microsoft Exchange is the existing calendaring application, use TMSXE. TMSXE is installed on a standalone server, and TMSXN is installed on the Lotus Domino server. The integration software is installed separately from Cisco TMS and communicates with your calendaring server using HTTP or HTTPS.

Cisco recommends having your video conferencing resources (Cisco TelePresence Video Communication Server or Cisco MCU) dedicated for either scheduled or permanent/instant conferences. This is because permanent or instant conferences could consume scheduled resources, which would result in undesirable consequences on the scheduled conferences, such as scheduled video participants being unable to join or joining as audio-only due to lack of resources on the server.

Reporting

Cisco TMS provides various types of reporting and analysis functionality, including:

- Asset management reports: ticket logs, device events, device alarms, and connectivity
- Detailed call history reports for managed endpoints and infrastructure
- Scheduling activity reports, including user-based, scheduling interface used, conference event logs, and conference reports

However, some of these functions work only in certain deployments. For example, when an endpoint such as the Cisco TelePresence TX9000 or Cisco TelePresence System EX90 is registered to Cisco Unified Communications Manager (Unified CM), Cisco TMS cannot generate reports for that endpoint. Cisco TMS can generate only call history and call detail record (CDR) reports for an endpoint registered to the Cisco TelePresence Video Communication Server (VCS). For those endpoints that are registered to Unified CM, CDRs can be downloaded from Unified CM.

Organizations that require more customized reports, business knowledge, and integration with Business Intelligence Applications can use the Cisco TelePresence Management Suite Analytics Extension (TMSAE), which is an online analytical processing system for Cisco TMS that provides advanced reporting functionality for your video network. For more information on Cisco TMSAE, refer to the product documentation available at

https://www.cisco.com/c/en/us/support/conferencing/telepresence-management-suite-extensions/tsd-products-support-series-home.html

Management

The main functions of management in the TelePresence environment include: provisioning, monitoring, maintenance, and resource management. Cisco TelePresence Management Suite (Cisco TMS) enables management of the TelePresence environment, along with a scheduling interface that it supports in a TelePresence environment.

Endpoint and Infrastructure Management

Cisco TMS can manage endpoints registered to both Cisco VCS and Cisco Unified CM. There are two types of device management: direct managed and provisioned.

Direct-managed devices are manually added into the Cisco TMS system navigator. Cisco TMS supports 5,000 direct-managed devices. Cisco TMS communicates with the endpoints directly via HTTP or SNMP protocols. When a direct-managed endpoint is registered to Unified CM, Unified CM handles most management capabilities such as software upgrades. When a direct-managed endpoint is registered to Cisco VCS, Cisco TMS handles management and provisioning of the endpoint, including capabilities such as software upgrades.

Cisco TMS can also directly manage infrastructure devices such as Cisco VCS, Cisco MCU's, and others. Currently Cisco TMS supports scheduling and management of conferencing devices registered with Cisco VCS only.

Provisioned endpoints are not in the TMS system navigator, but rather are provisioned by Cisco TMS through the Cisco TMS Provisioning Extension (TMSPE). Cisco TMS supports 100,000 provisioned devices. The provisioning method dramatically increases the scale that Cisco TMS can support. It also simplifies the procedure for a bulk deployment because there is no need to add the systems manually. However, Cisco TMS has less control on provisioned endpoints compared to direct-managed endpoints. In addition, scheduling is not supported for provisioned endpoints.

Cisco TMS provides the functionality of phone books for direct-managed endpoints registered to Cisco VCS as well as provisioned endpoints. Phone books provide ease of locating users and dialing out.

Cisco TMS also provides interfaces to monitor both scheduled and instant video conferences.

For more information, refer the latest version of the *Cisco TelePresence Management Suite Administrator Guide*, available at

https://www.cisco.com/c/en/us/support/conferencing/telepresence-management-suite-tms/products-maintenance-guides-list.html

Provisioning

The Cisco TMS Provisioning Extension (TMSPE) is a provisioning application for Cisco TMS and Cisco VCS. Cisco TMSPE enables video conferencing network administrators to create and manage large deployable video conferencing solutions. Cisco TMSPE is an add-on replacement for the TMS agent on the Cisco TMS server, and it provides the following main features:

- Ability to import users from Microsoft and generic LDAP sources (LDAP, LDAPS, AD)
- User personalization and administrative device configuration control for devices supported by Cisco TMS Provisioning Extension (for example, Jabber video, Cisco IP Video Phone E20, and Cisco TelePresence System EX Series and MX Series)
- Multi-tiered phone books for devices supported by Cisco TMSPE
- End-user FindMe portal on Cisco TMS using Microsoft Active Directory (AD) login instead of Cisco VCS Web user interface
- Support for up to 100,0000 users and devices

For further information, refer to the latest version of the *Cisco TelePresence Management Suite Provisioning Extension Deployment Guide*, available at

https://www.cisco.com/c/en/us/support/conferencing/telepresence-management-suite-tms/products-installation-and-configuration-guides-list.html

Phone books

Phone books help users maintain their contacts and dial them. Cisco TMS phone books can be created and populated from different sources such as Microsoft Active Directory (AD), Cisco Unified CM, an H.350 server, and gatekeepers.

There are two types of phone books: local phone books and global phone books. Local phone books (also called *favorites*) are a file stored on an endpoint specific to the end user. Contacts can be added, modified, and deleted as desired by the user.

Global or corporate phone books are pushed from Cisco TMS to the endpoints. They cannot be modified from the endpoint because they are automatically populated from AD, an H.350 server, or the local Cisco TMS database. Administrators can select the phone books for specific users and push them to the appropriate endpoints.

Maintenance and Monitoring

Cisco TMS has a Software Manager repository where the software images of endpoints and infrastructure devices can be added and then used to upgrade matching endpoints and infrastructure devices registered with Cisco VCS. Administrators can select a number of devices and upgrade them from Cisco TMS at one time. Cisco TMS provides the status of the upgrade. Using Cisco TMS to do upgrades is more convenient and easier than upgrading the endpoints and infrastructure devices manually.

Cisco TMS also provides monitoring capabilities for conferences. Cisco TMS lists all scheduled conferences, and the state of the conference (for example, Active) is displayed in TMS's Conference Control Center, with details of packet loss per participant for active conferences. Errors are displayed in TMS's Ticketing Service; for example, if there are configuration errors, Cisco TMS discovers them and opens a ticket associated with the appropriate device. Each ticket has an ID and severity level.

Cisco Smart Software Licensing

Cisco Collaboration System Release (CSR) 12.0 and later releases incorporate Cisco Smart Software Licensing and the Cisco Smart Software Manager (SSM) for management of an organization’s collaboration licenses. Cisco SSM provides a centralized method for applying, tracking, and managing licenses on Cisco Unified CM, Cisco Unity Connection, and Cisco Emergency Responder as well as other Cisco products. Cisco Smart Software Manager assists the administrator by automating many of the steps necessary to license users on the application servers.

Cisco Smart Software Licensing consists of the Cisco hosted Cisco Smart Software Manager web portal, where an organization's collaboration application entitlements and licenses are tracked and synchronized to collaboration components.

Customers purchase licenses, and these licenses are automatically applied to the customer's Cisco Smart Account and synchronized via Cisco SSM with the on-premises applications. Cisco Smart Software Manager registers on-premises collaboration application instances to Cisco Licensing Services and synchronizes the organization’s licenses against the applications.

The following Unified Communications applications use Cisco Smart Software Licensing:

- Cisco Unified Communications Manager (Unified CM) — including Cisco IM and Presence, which is licensed through Unified CM, and Cisco Unified Communications Manager Session Management Edition (Unified CM SME)
- Cisco Unity Connection
- Cisco Emergency Responder

Appropriate licenses must first be acquired and applied to the Cisco Smart Account for managing software and entitlement using the Cisco Smart Software Manager portal. Next, an organization administrator generates a product instance registration token on the Cisco Smart Software Manager portal at https://software.cisco.com. The administrator then registers the collaboration application product instance using the registration token copied from the Cisco Smart Software Manager portal. Once registered, the applications will synchronize with Cisco Smart Software Manager and receive user and feature licensing entitlement information.

An application is allowed 90 days of non-compliance, during which the system will function normally and the administrators can make changes if there are insufficient licenses or if the system has lost communication with Cisco SSM. If the system remains out of compliance (that is, if sufficient licenses are not acquired or communication to Cisco SSM is not restored) for 90 days, the collaboration application functionality is reduced as follows:

- Cisco Unified Communications Manager (Unified CM) — call control

  When the system is out of compliance for 90 days, Unified CM will continue to handle calls but no user or device moves, adds, changes, or deletions (MACD) will be allowed.

- Cisco Unity Connection — voice messaging

  When the system is out of compliance, the system will continue to allow administrative changes but will not provide voice messaging services. That is, the system will no longer answer calls, thus preventing callers from leaving messages and preventing users from retrieving voice messages.

- Cisco Emergency Responder

  When the system is out of compliance, the Cisco Phone Tracking Engine service is stopped and the system stops tracking phones and updating locations.

For more information on Cisco Smart Software Licensing and licensing management with the Cisco Smart Software Manager, refer to the information at

https://www.cisco.com/go/smartlicensing

Deployment Scenarios

Cisco Smart Software Licensing is automatically enabled on the publisher node of all supported applications. Licensing is managed by the Cisco Smart Licensing Manager Service, which is activated and started automatically on the publisher node in each cluster. The publisher node manages the licensing for all other nodes in the cluster.

In order for the collaboration applications to register and synchronize licensing information with the Cisco Smart Software Manager (SSM), the Cisco Smart Licensing Manager Service running on the publisher node of each cluster must communicate over the Internet to Cisco SSM services. This communication is either direct or is mediated by an intermediary.

The collaboration application attempts to communicate directly with the Cisco SSM service over the Internet using HTTPS. In some organizations, outbound HTTPS traffic is allowed, and this traffic is passed to the online Cisco SSM service without issue. In cases where organizations do not allow direct outbound HTTPS traffic from their data center applications to the Internet, Cisco Smart Software Licensing communications may be directed to the Internet by an HTTP proxy. In either case, when the application communicates with the Cisco SSM online services (with or without an HTTP proxy), it does so directly over the Internet.

Alternatively, communications between the collaboration application publisher nodes may be directed to an on-premises Cisco Smart Software Manager satellite system. This is the mediated method of Smart Licensing communications. The Cisco SSM satellite system is a virtual machine (VM) deployed in the on-premises data center. The SSM satellite system acts as an intermediary and relays communications between the on-premises collaboration applications and the Internet-hosted, online Cisco SSM service. The SSM satellite must periodically connect to the online Cisco SSM service to synchronize. This periodic synchronization is facilitated by direct HTTPS communications between the Cisco SSM satellite system and the online Cisco SSM service. This is the Cisco SSM satellite connected mode. As mentioned previously, if the organization has restrictions on outbound HTTPS traffic to the Internet, either an HTTP proxy is used or, alternatively, a report file from the SSM satellite system may be manually uploaded periodically to the online service to maintain registration and authorization.

The main consideration for choosing between direct and mediated deployments of Cisco SSM is the organization's network and security policies related to Internet and online services access. If your organization restricts outbound access to the Internet, consider mediated deployments with Cisco SSM satellite, noting the requirement for a separate Cisco Smart Software Manager satellite VM in the on-premises data center.

Deployment Recommendations

Generally speaking, direct or proxy communication is recommended between the on-premises collaboration application (Unified CM, Unity Connection, and Emergency Responder) cluster publisher nodes and the web-hosted Cisco Smart Software Manager service. This does require outbound HTTPS communications from the application publisher nodes through the organization's firewall to the Cisco Smart Software Manager service. If the organization's policy does not allow for direct outbound web communications, cluster publisher nodes may leverage a new or existing standard HTTP/HTTPS proxy server within the organization to enable firewall traversal and access to the web-hosted Cisco Smart Software Manager service.

When creating product registration tokens and registering collaboration application publisher instances within Cisco SSM, administrators may use different virtual accounts and/or registration tokens under the same Smart Account to group specific product types or products in specific locations for ease of license administration and management.

Creating multiple virtual accounts is recommended because it simplifies license management by allowing licenses to be pooled and shared across multiple products and virtual accounts across the organization. Further, by segmenting product instances and tokens across virtual accounts, organizations can more easily track and account for the costs of licenses along more granular accounting lines in order to better manage operating costs and other expenses.

Note Because licenses may not be pooled, moved, or managed across different Cisco Smart Accounts, Cisco recommends that the organization establish only a single Smart Account (with multiple Virtual Accounts as required) unless there are specific organizational policies or requirements (regulations, laws, or other restrictions) that offset the limitations of multiple Smart Accounts.

Redundancy

The online Cisco SSM service is highly available; however, in the case of an Internet connection issue where the collaboration application system is out of compliance, the system will continue to operate normally for 90 days. User and device provisioning is not possible once the system reaches full non-compliance. In order to maintain normal system operation, the online Cisco SSM must be reachable consistently.

When Cisco Smart Software Licensing is deployed in mediated mode with the Cisco SSM satellite system, install and configure at least two SSM satellite VMs to ensure high availability. Cisco SSM satellite systems rely on an active/standby redundancy scheme, whereby if the active or primary system fails or loses connectivity to collaboration applications or the online Cisco SSM service, the standby or backup system takes over SSM operations.

Capacity Planning for Cisco Smart Software Manager

The online Cisco SSM service provides near infinite scale, given the elastic nature of compute resources within the service data centers. This means that, effectively, an organization can license infinite numbers of collaboration applications using Cisco SSM.

On the other hand, when running in mediated mode, Cisco SSM satellite VMs do have a capacity limit. A single Cisco SSM satellite VM can handle up to 4,000 product instance registrations. When deploying Cisco SSM in mediated mode, be sure to deploy VMs in sufficient quantity to handle all the product instances that require licenses for the deployment.

Additional Tools

In addition to the network management tools mentioned above, the following tools also provide troubleshooting and reporting capabilities for Cisco Unified Communications systems:

- Cisco Unified Analysis Manager, page 27-24
- Cisco Unified Reporting, page 27-25

Cisco Unified Analysis Manager

Cisco Unified Analysis Manager is included with the Cisco Unified Communications Manager Real-Time Monitoring Tool (RTMT). RTMT runs as a client-side application and it uses HTTPS and TCP to monitor system performance, device status, device discovery, and CTI applications for Unified CM. RTMT can connect directly to devices via HTTPS to troubleshoot system problems.

Unlike the other RTMT functions, Unified Analysis Manager is unique in that it supports multiple Unified Communications elements instead of just one. When the Unified Analysis Manager is launched, it collects troubleshooting information from your Unified Communications system and provides an analysis of that information. You can use this information to perform your own troubleshooting operations, or you can send the information to Cisco Technical Assistance Center (TAC) for analysis.

Unified Analysis Manager supports the following Unified Communications elements:

- Cisco Unified Communications Manager
- Cisco Unified Contact Center Enterprise
- Cisco Unified Contact Center Express
- Cisco IOS Voice Gateways
- Cisco Unity Connection
- Cisco IM and Presence

Unified Analysis Manager provides the following key features and capabilities:

- Supports collection of Unified Communications application hardware, software, and license information from Unified Communications elements.
- Supports setting and resetting of trace level across Unified Communications elements.
- Supports collection and export to a define FTP server of log and trace files from Unified Communications elements.
- Supports analysis of the call path (call trace capability) across Unified Communications elements.

For more details on the report options, refer to the information about the Cisco Unified Analysis Manager in the latest version of the *Cisco Unified Real-Time Monitoring Tool Administration Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

Cisco Unified Reporting

The Cisco Unified Reporting web application generates reports for troubleshooting or inspecting Cisco Unified Communications Manager cluster data. It is a convenient tool that you can access from the Unified Communications Manager console. The tool facilitates gathering data from existing sources, comparing the data, and reporting irregularities. For example, you can view a report that shows the hosts file for all servers in the cluster. The application gathers information from the publisher server and each subscriber server. Each report provides data for all active cluster nodes that are accessible at the time the report is generated.

For example, the following reports can be used for general management of a Unified CM cluster:

- Unified CM Cluster Overview — Provides an overview of the cluster, including Unified CM version, hostname, and IP address of all servers, a summary of the hardware details, and so forth.
- Unified CM Device Counts Summary — Provides the number of devices by model and protocol that exist in the Cisco Unified Communications Manager database.

The following report can be used for debugging a Unified CM cluster:

- Unified CM Database Replication Debug — Provides debugging information for database replication.

The following report can be used for maintenance of a Unified CM cluster:

- Unified CM Database Status - Provides a snapshot of the health of the Unified CM database. This report should be generated before an upgrade to ensure the database is healthy.

For more information on the report options, refer to the latest version of the *Cisco Unified Reporting Administration Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html

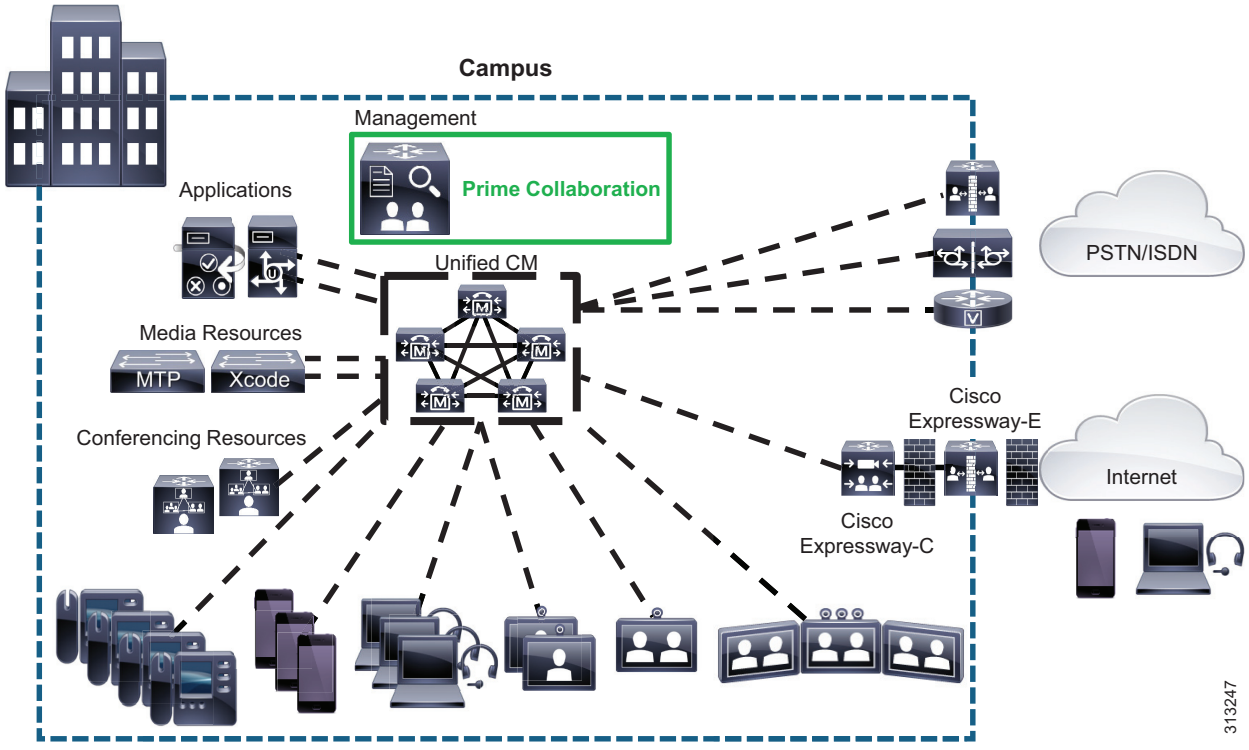
Integration with Cisco Collaboration Deployment Models

This section discusses how to deploy Cisco Collaboration and network management applications in various deployment models. For detailed information on the deployment models, see the chapter on Collaboration Deployment Models, page 10-1.

Campus

In the campus model, Cisco network management applications, along with call processing agents, are deployed at a single site (or campus) with no telephony services provided over an IP WAN. An enterprise would typically deploy the single-site model over a LAN or metropolitan area network (MAN). Figure 27-2 illustrates the deployment of Cisco network management applications in the single-site model.

Figure 27-2 Campus Deployment



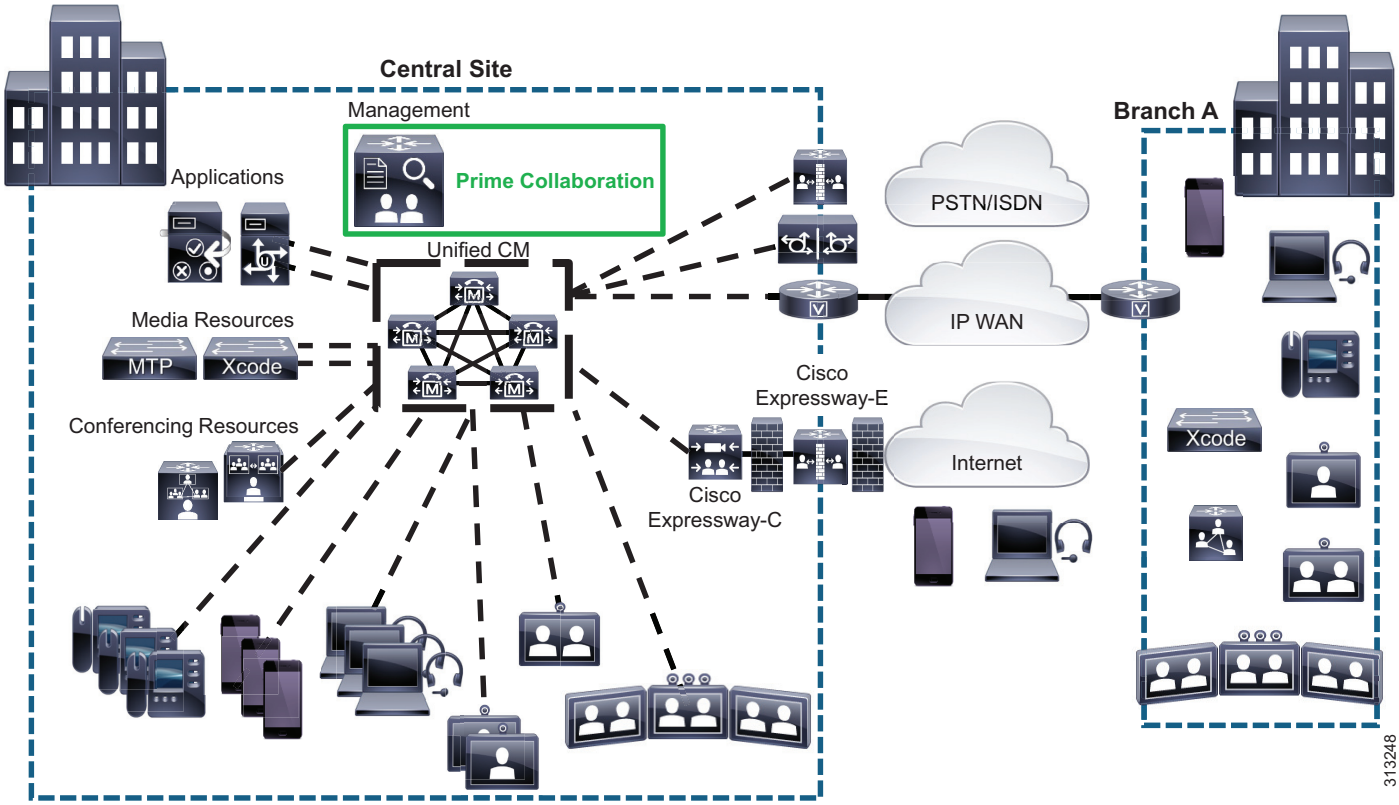
The following design characteristics and recommendations apply to the single-site model for deploying Prime Collaboration:

- Cisco recommends deploying Unified CM voice quality monitoring to monitor overall voice quality in the network.
- Cisco recommends deploying the Cisco NAM to monitor key IP phone devices, gateway devices, and application servers in the network and to investigate and troubleshoot voice quality issues.

Multisite WAN with Centralized Call Processing

The multisite WAN model with centralized call processing is really an extension of single-site model, with an IP WAN between the central site and remote sites. The IP WAN is used to transport voice traffic between the sites and call control signaling between the central site and the remote sites. Figure 27-3 illustrates the deployment of Cisco network management applications in a multisite WAN model with centralized call processing.

Figure 27-3 Multisite WAN Deployment with Centralized Call Processing



The following design characteristics and recommendations apply to the multisite model for deploying Prime Collaboration with centralized call processing:

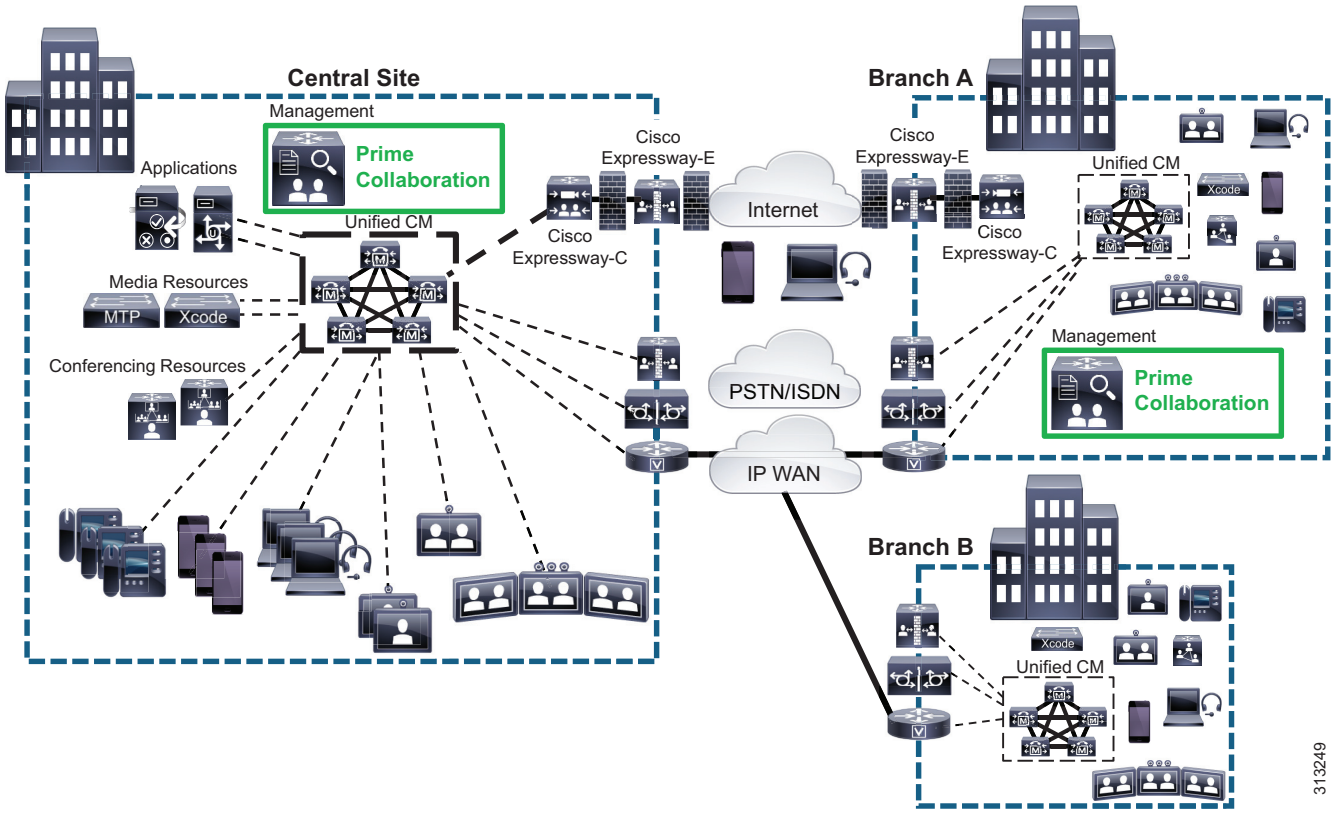
- Cisco recommends deploying all network management applications (including Prime Collaboration) in the central site to locate them with the call processing agent. The benefit of such an implementation is that it keeps the network management traffic between call processing agent and network management applications within the LAN instead of sending that traffic over the WAN circuit.
- Multiple Prime Collaborations can be deployed, with each instance managing multi-site and multi-cluster Unified Communications environments. In this deployment scenario, Cisco recommends that you deploy a Manager of Managers (MoM). Each Prime Collaboration can provide real-time notifications to the higher-level MoM using SNMP traps, syslog notifications, and email to report the status of the network being monitored.
- Cisco recommends deploying Unified CM voice quality monitoring to monitor overall voice quality in the network.

- Cisco recommends using the Service Level Agreement (SLA) feature and Synthetic test feature to check for network infrastructure status.
- Cisco recommends deploying the Cisco NAM to monitor key IP phone devices, gateway devices, and application servers in the network and to investigate and troubleshoot voice quality issues.

Multisite WAN with Distributed Call Processing

The multisite WAN model with distributed call processing consists of multiple independent sites, each with its own call processing agent connected to an IP WAN. Figure 27-4 illustrates the deployment of Cisco network management applications in a multisite WAN model with distributed call processing.

Figure 27-4 Multisite WAN Deployment with Distributed Call Processing



A multisite WAN deployment with distributed call processing has many of the same requirements as a single site or a multisite WAN deployment with centralized call processing in terms of deploying Prime Collaboration. Follow the best practices and recommendations from these other models in addition to the ones listed here for the distributed call processing model:

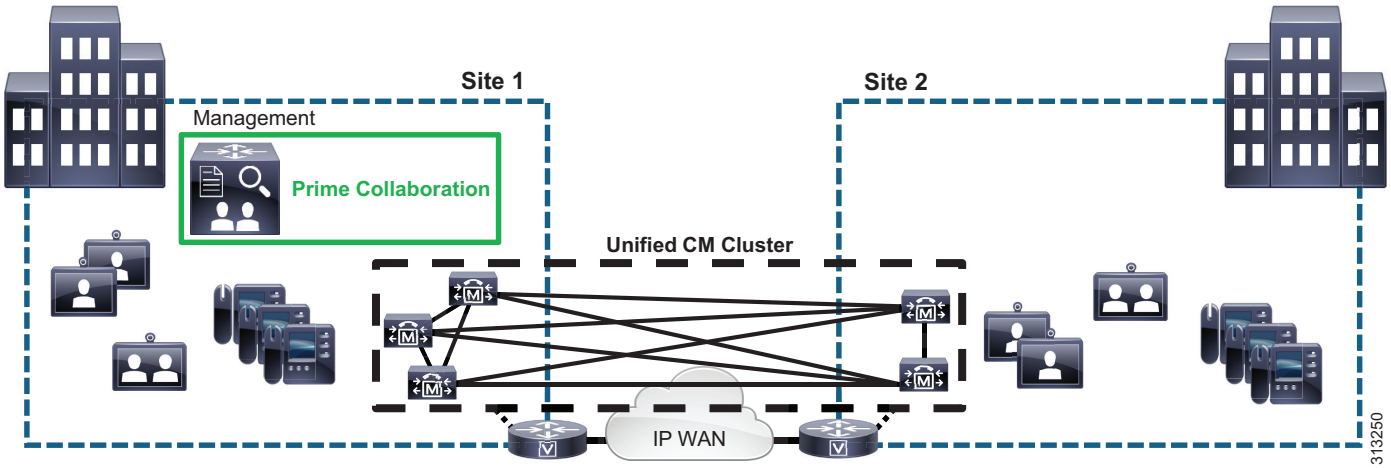
- If only one Cisco Prime Collaboration deployment is used to manage multiple Unified CM clusters, Cisco recommends deploying Prime Collaboration along with the Unified CM cluster that has the highest call volume and the most endpoints.

- Multiple Prime Collaborations can be deployed, with each instance managing multi-site and multi-cluster Unified Communications environments. In this deployment scenario, Cisco recommends that you deploy a Manager of Managers (MoM). Each Prime Collaboration can provide real-time notifications to the higher-level MoM using SNMP traps, syslog notifications, and email to report the status of the network being monitored.
- Cisco recommends deploying Unified CM voice quality monitoring to monitor overall voice quality in the network.
- Cisco recommends deploying the Cisco NAM to monitor key IP phone devices, gateway devices, and application servers in the network and to investigate and troubleshoot voice quality issues.

Clustering over the WAN

Clustering over the WAN refers to a single Cisco Unified CM cluster deployed across multiple sites that are connected by an IP WAN with QoS features enabled. This deployment model is designed to provide call processing resiliency if the IP WAN link fails. Figure 27-5 illustrates the deployment of Cisco network management applications with clustering over the WAN.

Figure 27-5 Clustering over the WAN



Note There is no native high-availability or redundancy support for Prime Collaboration with this model.

The following design characteristics and recommendations apply when deploying Prime Collaboration with clustering over the WAN:

- Cisco recommends deploying Prime Collaboration in the headquarter site where Unified CM publisher is located.
- Multiple Prime Collaborations can be deployed, with each instance managing multi-site and multi-cluster Unified Communications environments. In this deployment scenario, Cisco recommends that you deploy a Manager of Managers (MoM). Each Prime Collaboration can provide real-time notifications to the higher-level MoM using SNMP traps, syslog notifications, and email to report the status of the network being monitored.

- Cisco recommends deploying Unified CM voice quality monitoring to monitor overall voice quality in the network.
- Cisco recommends deploying the Cisco NAM to monitor key IP phone devices, gateway devices, and application servers in the network and to investigate and troubleshoot voice quality issues.

Revised: March 1, 2018

A

AA	Automated attendant
AAD	Alerts and Activities Display
AAR	Automated Alternate Routing
AC	Cisco Attendant Console
ACD	Automatic call distribution
ACE	Cisco Application Control Engine
ACF	Admission Confirm
ACL	Access control list
ACS	Access Control Server
AD	Microsoft Active Directory
ADAM	Active Directory Application Mode
ADFS	Microsoft Active Directory Federated Services
ADPCM	Adaptive Differential Pulse Code Modulation
ADUC	Active Directory Users and Computers
AES	Advanced Encryption Standards
AFT	ALI Formatting Tool
AGM	Cisco Access Gateway Module
ALG	Application Layer Gateway
ALI	Automatic Location Identification
AMI	Alternate mark inversion
AMIS	Audio Messaging Interchange Specification
AMWI	Audible message waiting indication

ANI	Automatic Number Identification
AP	Access point
APDU	Application protocol data unit
API	Application Program Interface
APNs	Apple Push Notification service
ARJ	Admission Reject
ARP	Address Resolution Protocol
ARQ	Admission Request
ASA	Cisco Adaptive Security Appliance
ASP	Active server page
ASR	Automatic speech recognition
ATA	Cisco Analog Telephone Adaptor
ATM	Asynchronous Transfer Mode
AVC	Advanced Video Coding
AXL	Administrative XML Layer

B

BAT	Cisco Bulk Administration Tool
BBWC	Battery-backed write cache
BES	Blackberry Enterprise Server
BFCP	Binary Floor Control Protocol
BFD	Bidirectional Forwarding Detection
BGP	Border Gateway Protocol
BHCA	Busy hour call attempts
BHCC	Busy hour call completions
BIB	Built-in bridge
BLF	Busy lamp field
BOSH	Bidirectional-streams Over Synchronous HTTP

BPDU	Bridge protocol data unit
bps	Bits per second
BRI	Basic Rate Interface
BTN	Bill-to number

C

CA	Certificate Authority
CAC	Call admission control
CAM	Content-addressable memory
CAMA	Centralized Automatic Message Accounting
CAPF	Certificate Authority Proxy Function
CAPWAP	Control and Provisioning of Wireless Access Points
CAR	Cisco CDR Analysis and Reporting
CAS	Channel Associated Signaling
CBWFQ	Class-Based Weighted Fair Queuing
CCA	Clear channel assessment
CCD	Call Control Discovery
CCS	Common channel signaling
CDI	Cisco Directory Integration
CDP	Cisco Discovery Protocol
CDR	Call detail record
CER	Cisco Emergency Responder
CGI	Common Gateway Interface
CIF	Common Intermediate Format
CIR	Committed information rate
CKM	Cisco Centralized Key Management
CLEC	Competitive local exchange carrier
CLID	Calling line identifier

CM	Cisco Unified Communications Manager (Unified CM)
CMC	Client Matter Code
CME	Cisco Unified Communications Manager Express (Unified CME)
CMI	Cisco Messaging Interface
CMM	Cisco Communication Media Module
CMR	Call management record
CMR	Collaboration Meeting Room
CO	Central office
Co-located	Two or more devices in the same physical location, with no WAN or MAN connection between them
Co-resident	Two or more services or applications running on the same server or virtual machine
COM	Component Object Model
COP	Cisco Options Package
COR	Class of restriction
CoS	Class of service
CPCA	Cisco Unity Personal Assistant
CPI	Cisco Product Identification tool
CPL	Call Processing Language
CPN	Calling party number
CRM	Customer relationship management
CRS	Cisco Customer Response Solution
cRTP	Compressed Real-Time Transport Protocol
CSF	Client Services Framework
CSTA	Computer-Supported Telecommunications Applications
CSUF	Cross-Stack UplinkFast
CSV	Comma-separated values
CTI	Computer telephony integration
CTL	Certificate Trust List
CUBE	Cisco Unified Border Element, formerly the Cisco Multiservice IP-to-IP Gateway (IP-IP Gateway)
CUE	Cisco Unity Express

CUMI	Cisco Unity Connection Messaging Interface
CUPI	Cisco Unity Connection Provisioning Interface
CUSP	Cisco Unified SIP Proxy
CVTQ	Cisco Voice Transmission Quality
CWA	Microsoft Office Communicator Web Access

D

DC	Domain controller
DDNS	Dynamic Domain Name Server
DDR	Delayed Delivery Record
DECT	Digital Equipment Cordless Telephony
DFS	Dynamic Frequency Selection
DHCP	Dynamic Host Configuration Protocol
DID	Direct inward dial
DIT	Directory Information Tree
DMVPN	Dynamic Multipoint Virtual Private Network
DMZ	Demilitarized zone
DN	Directory number
DNIS	Dialed number identification service
DNS	Domain Name System
DoS	Denial of service
DPA	Digital PBX Adapter
DRS	Disaster Recovery System
DSCP	Differentiated Services Code Point
DSE	Digital set emulation
DSP	Digital signal processor
DTIM	Delivery Traffic Indicator Message
DTLS	Datagram Transport Layer Security protocol

DTMF Dual tone multifrequency

DTPC Dynamic Transmit Power Control

DUC Domino Unified Communications Services

E

E-SRST Enhanced Survivable Remote Site Telephony

E&M Receive and transmit, or ear and mouth

EAP Extensible Authentication Protocol

EAPOL Extensible Authentication Protocol over LAN

EC Echo cancellation

ECC Extended Call Context

ECDSA Elliptical Curve Digital Signature Algorithm

ECM Error correction mode

ECS Empty Capabilities Set

EI Enhanced Image

EIGRP Enhanced Interior Gateway Routing Protocol

ELCAC Enhanced Location Call Admission Control

ELIN Emergency location identification number

EM Extension Mobility

EMCC Extension Mobility Cross Cluster

ER Cisco Emergency Responder

ERL Emergency response location

ESF Extended Super Frame

F

FAC Forced Authorization Code

FCC Federal Communications Commission

FCoE Fibre Channel over Ethernet

FECC	Far End Camera Control
FIFO	First-in, first-out
FQDN	Fully qualified domain name
FR	Frame Relay
FTP	File Transfer Protocol
FWSM	Firewall Services Module
FXO	Foreign Exchange Office
FXS	Foreign Exchange Station

G

GARP	Gratuitous Address Resolution Protocol
GC	Global catalog
GDPR	Global Dial Plan Replication
GKTMP	Gatekeeper Transaction Message Protocol
GLBP	Gateway Load Balancing Protocol
GMS	Greeting management system
GPO	Group Policy Object
GPRS	General Packet Radio Service
GSB	Global Site Backup
GSM	Global System for Mobile Communication
GSS	Global Site Selector
GUI	Graphical user interface
GUP	Gatekeeper Update Protocol

H

H.225D	H.225 daemon
HDLC	High-Level Data Link Control
HMS	Hardware Media Server

HP Hewlett-Packard

HSRP Hot Standby Router Protocol

HTTP Hyper-Text Transfer Protocol

HTTPS HTTP Secure

HVD Hosted virtual desktop

Hz Hertz

I

IANA Internet Assigned Numbers Authority

IAPP Inter-Access Point Protocol

ICA Independent Computing Architecture

ICCS Intra-Cluster Communication Signaling

ICE Interactive Connectivity Establishment

ICMP Internet Control Message Protocol

ICS IBM Cabling System

ICT Intercluster trunk

IdP Identity Provider

IE Information Element

IETF Internet Engineering Task Force

IGMP Internet Group Management Protocol

IIS Microsoft Internet Information Server

ILS Intercluster Lookup Service

IM Instant messaging

IMAP Internet Message Access Protocol

IMS IP Multimedia Subsystem

IntServ Integrated Services

IntServ/DiffServ Integrated Services/Differentiated Services

IOPS Input/output operations per second

IP	Internet Protocol
IPCC	Cisco IP Contact Center
IPMA	Cisco IP Manager Assistant
IPPM	Cisco IP Phone Messenger
IPSec	IP Security
IP SLA VO	IP Service Level Agreement Video Operation
IPVMS	Cisco IP Voice Media Streaming Application
ISO	International Standards Organization
ISR	Integrated Services Router
ITEM	CiscoWorks IP Telephony Environment Monitor
ITL	Initial Trust List
ITU	International Telecommunication Union
IVR	Interactive voice response

J

JCF	Jabber Client Framework
JID	Jabber Identifier
JTAPI	Java Telephony Application Programming Interface

K

kbps	Kilobits per second
KEM	Key expansion module
KPML	Key Press Markup Language

L

LAN	Local area network
LBM	Location Bandwidth Manager
LBR	Low bit-rate

LCD	Liquid crystal display
LCF	Location Confirm
LCS	Live Communications Server
LDAP	Lightweight Directory Access Protocol
LDAPS	LDAP over SSL
LDIF	LDAP Data Interchange Format
LDN	Listed directory number
LEAP	Lightweight Extensible Authentication Protocol
LEC	Local Exchange Carrier
LFI	Link fragmentation and interleaving
LHS	Left-hand side
LLDP	Link Layer Discovery Protocol
LLDP-MED	Link Layer Discovery Protocol for Media Endpoint Devices
LLQ	Low-latency queuing
LRG	Local route group
LRJ	Location Reject
LRQ	Location Request
LSC	Locally significant certificate
LUN	Logical unit number
LWAP	Light Weight Access Point
LWAPP	Light Weight Access Point Protocol

M

MAC	Media Access Control
MAN	Metropolitan area network
Mbps	Megabits per second
MCM	Multimedia Conference Manager
MCS	Media Convergence Server

MCU	Multipoint Control Unit
MDN	Mobile Data Network
MDS	Mobile Data Services
MFT	Multiflex trunk
MGCP	Media Gateway Control Protocol
MIB	Management Information Base
MIC	Manufacturing installed certificate
MIME	Multipurpose Internet Mail Extension
MIPS	Millions of instructions per second
MISTP	Multiple Instance Spanning Tree Protocol
MITM	Man-in-the-middle
MLA	Cisco Multi-Level Administration
MLP	Multilink Point-to-Point Protocol
MLPP	Multilevel Precedence and Preemption
MLPPP	Multilink Point-to-Point Protocol
MLTS	Multi-line telephone system
MMoIP	Multimedia Mail over IP
MMP	Mobile Multiplexing Protocol
MOC	Microsoft Office Communicator
MoH	Music on hold
MOS	Mean Opinion Score
MPLS	Multiprotocol Label Switching
MRD	Media Routing Domain
MRG	Media resource group
MRGL	Media resource group list
ms	Millisecond
MSI	Media Services Interface
MSP	Managed service provider
MSP	Media Services Proxy

MTLS Mutual Transport Layer Security

MTP Media termination point

mW Milli-Watt

MWI Message Waiting Indicator

MXE Media Experience Engine

N

NAS Network Attached Storage

NAT Network Address Translation

NDR Non-delivery receipt

NENA National Emergency Number Association

NFAS Non-Facility Associated Signaling

NIC Network interface card

NOC Network operations center

NPA Numbering Plan Area

NSE Named Service Event

NSF Network Specific Facilities

NTE Named Telephony Event

NTLP Network Transmission Loss Plan

NTP Network Time Protocol

O

OBTP One Button To Push

OCS Microsoft Office Communicator Server

ORA Open Recording Architecture

OSPF Open Shortest Path First

OU Organizational unit

OVA	Open Virtualization Archive
OWA	Outlook Web Access

P

PAC	Protected Access Credential
PAK	Product Activation Key
PBX	Private branch exchange
PC	Personal computer
PCAP	Phone Control and Presence
PCI	Peripheral Component Interconnect
PCM	Pulse code modulation
PCoIP	PC over IP
PCTR	Personal call transfer rule
PD	Powered device
PHB	Per-hop behavior
PII	Personally Identifiable Information
PIN	Personal identification number
PINX	Private integrated services network exchange
PIX	Private Internet Exchange
PKI	Public Key Infrastructure
PLAR	Private Line Automatic Ringdown
POD	Piece of Data
PoE	Power over Ethernet
POTS	Plain old telephone service
PPP	Point-to-Point Protocol
pps	Packets per second
PQ	Priority Queue
PRACK	Provisional Reliable Acknowledgement

PRI	Primary Rate Interface
PSAP	Public safety answering point
PSE	Power source equipment
PSK	Pre-Shared Key
PSTN	Public switched telephone network
PVC	Permanent virtual circuit

Q

QBE	Quick Buffer Encoding
QBSS	QoS Basic Service Set
QoS	Quality of Service
QSIG	Q signaling

R

RADIUS	Remote Authentication Dial-In User Service
RAS	Registration Admission Status
RBAC	Role-Based Access Control
RCC	Remote Call Control
RCP	Remote Copy Protocol
RDNIS	Redirected Dialed Number Information Service
REST	Representational State Transfer
RF	Radio frequency
RFC	Request for Comments
RHS	Right-hand side
RIM	Research In Motion
RIP	Routing Information Protocol
RIS	Real-Time Information Server

RMA	Remote Mobile Access
RMTP	Reliable Multicast Transport Protocol
RoST	RSVP over SIP Trunks
RSA	Rivest, Shamir, and Adelman
RSNA	Reservationless Single Number Access
RSP	Route/Switch Processor
RSSI	Relative Signal Strength Indicator
RSTP	Rapid Spanning Tree Protocol
RSVP	Resource Reservation Protocol
RTCP	Real-Time Transport Control Protocol
RTMP	Real-Time Messaging Protocol
RTMT	Cisco Real-Time Monitoring Tool
RTP	Real-Time Transport Protocol
RTSP	Real Time Streaming Protocol
RTT	Round-trip time

S

S1, S2, S3, and S4	Severity levels for service requests
SaaS	Software-as-a-Service
SAF	Service Advertisement Framework
SAML	Security Assertion Markup Language
SAN	Storage area networking
SBC	Session Border Controller
SCCP	Skinny Client Control Protocol
SCSI	Small Computer System Interface
SDI	System Diagnostic Interface
SDK	Software Development Kit

SDL	Signaling Distribution Layer
SDP	Session Description Protocol
SE	Cisco Systems Engineer
SF	Super Frame
SFTP	Secure File Transfer Protocol
SI	Standard Image
SIMPLE	SIP for Instant Messaging and Presence Leveraging Extensions
SIP	Session Initiation Protocol
SIS	Symbian installation system
SIW	Service Inter-Working
SLA	Service level agreement
SLA VO	IP Service Level Agreement Video Operation
SLB	Server load balancing
SLDAP	Secure LDAP
SMA	Segmented Meeting Access
SMDI	Simplified Message Desk Interface
SME	Cisco Unified Communications Manager Session Management Edition
SMS	Short Message Service
SMTP	Simple Mail Transfer Protocol
SNMP	Simple Network Management Protocol
SOAP	Simple Object Access Protocol
SPA	Shared Port Adapter
SPAN	Switched Port Analyzer
SQL	Structured Query Language
SRE	Cisco Services-Ready Engine
SRND	Solution Reference Network Design
SRST	Survivable Remote Site Telephony
SRSV	Survivable Remote Site Voicemail
SRTP	Secure Real-Time Transport Protocol

SRV	Server
SS7	Signaling System 7
SSID	Service set identifier
SSL	Secure Sockets Layer
SSM	Cisco Smart Software Manager
SSO	Single Sign-On
STP	Spanning Tree Protocol
STUN	Session Traversal Utilities for NAT
SUP1	Cisco Supervisor Engine 1
SUP2	Cisco Supervisor Engine 2
SUP2+	Cisco Supervisor Engine 2+
SUP3	Cisco Supervisor Engine 3

T

TAC	Cisco Technical Assistance Center
TAPI	Telephony Application Programming Interface
TCD	Telephony Call Dispatcher
TCER	Total Character Error Rate
TCL	Tool Command Language
TCP	Transmission Control Protocol
TCS	Terminal Capabilities Set
TDD	Telephone Device for the Deaf
TDM	Time-division multiplexing
TEHO	Tail-end hop-off
TFTP	Trivial File Transfer Protocol
TIP	Telepresence Ineroperability Protocol
TKIP	Temporal Key Integrity Protocol
TLS	Transport Layer Security

TMS	Cisco TelePresence Management Suite
TMSXE	Cisco TelePresence Management Suite Extension for Microsoft Exchange
ToD	Time of day
ToS	Type of service
TPC	Transmit Power Control
TRaP	Telephone record and playback
TRC	Tested Reference Configuration
TRP	Trusted Relay Point
TSP	Teleconferencing Service Provider
TTL	Time to live
TTS	Text-to-speech
TTY	Terminal teletype
TUI	Telephony user interface
TURN	Traversal Using Relays around NAT

U

UAC	User agent client
UAS	User agent server
UCCN	Unified Client Change Notifier
UCS	Cisco Unified Computing System
UDC	Universal data connector
UDLD	UniDirectional Link Detection
UDP	User Datagram Protocol
UDPTL	Unnumbered Datagram Protocol Transport Layer
UDS	User Data Service
UMTS	Universal Mobile Telecommunications System
UN	Unsolicited SIP Notify
UNC	Universal Naming Convention

UP	User Priority
UPS	Uninterrupted power supply
URI	Uniform resource identifier
USB	Universal Serial Bus
UTIM	Cisco Unity Telephony Integration Manager
UTP	Unshielded twisted pair
UUIE	User-to-User Information Element

V

V3PN	Cisco Voice and Video Enabled Virtual Private Network
VAD	Voice activity detection
VAF	Voice-Adaptive Fragmentation
VATS	Voice-Adaptive Traffic Shaping
VCS	Cisco TelePresence Video Communication Server
VDI	Virtual Desktop Infrastructure
VDS	VMware vSphere Distributed Switch
VIC	Voice interface card
VLAN	Virtual local area network
VMO	Cisco ViewMail for Outlook
VoIP	Voice over IP
VoPSTN	Voice over the PSTN
VoWLAN	Voice over Wireless LAN (WLAN)
VPIM	Voice Profile for Internet Mail protocol
VPN	Virtual private network
VRRP	Virtual Router Redundancy Protocol
VUI	Voice User Interface
VVB	Cisco Virtualized Voice Browser
VWIC	Voice/WAN interface card

VXC	Cisco Virtualization Experience Client
VXI	Cisco Virtualization Experience Infrastructure
VXME	Virtualization Experience Media Engine

W

WAN	Wide area network
WebDAV	Web-Based Distributed Authoring and Versioning
WEP	Wired Equivalent Privacy
WFQ	Weighted fair queuing
WINS	Windows Internet Naming Service
WLAN	Wireless local area network
WLC	Wireless LAN controller
WLSM	Cisco Wireless LAN Services Module
WMM	Wi-Fi Multimedia
WMM TSPEC	Wi-Fi Multimedia Traffic Specification
WPA	Wi-Fi Protected Access

X

XCP	Extensible Communications Platform
XML	Extensible Markup Language
XMPP	Extensible Messaging and Presence Protocol

Symbols

! in route patterns 14-27

@ in route patterns 14-26

+ dialing 14-57

+E.164 numbering plan 14-75

Numerics

3500 Series Video Gateways 5-11

3900 Series SIP Phones 8-10

508 conformance 8-5

7800 Series Phones 8-8

7900 Series Phones 8-8

7905_7912 dial rules 14-20

7921G Wireless IP Phone 8-33

7925G-EX Wireless IP Phone 8-33

7925G Wireless IP Phone 8-33

7926G Wireless IP Phone 8-33

7940_7960_OTHER dial rules 14-20

802.1s 3-4

802.1w 3-4, 3-7

802.1X authentication 4-12

802.3af PoE 3-12

8800 Series Phones 8-9, 8-15

9.@ route pattern 14-26, 14-27

911 calls 14-70, 15-1

A

AA 19-22

AAR

- dial plan considerations 14-70, 14-79
- for video calls 5-34
- for Voice over PSTN 10-22
- with Cisco Unity 19-7

AC 18-42

access codes 14-80, 21-54

access control list (ACL) 4-32

accessibility of endpoint features 8-5

Access Layer 3-4

access lists for Single Number Reach calls 21-57

access numbers 21-65

access point (AP) 3-61, 3-63, 3-72, 8-33, 15-11

access tokens 16-56

ACL 4-32

Active Directory (AD) 16-10, 16-15, 16-20, 16-26

Active Directory Application Mode (ADAM) 16-12, 16-31

Active Directory Lightweight Directory Services (AD LDS) 16-22

AD 16-10, 16-15, 16-20, 16-26

ADAM 16-12, 16-31

Adaptive Security Appliance (ASA) 4-33, 4-39

addresses

- flat 21-23
- MAC 4-7
- security 4-5
- security issues 4-4

Address Resolution Protocol (ARP) 3-72, 4-11

AD LDS 16-22

Administrative XML Layer (AXL) 27-7

advanced formulas for bandwidth calculations 3-59

AFT 15-29

agent desktop 23-10

agents for call processing 10-25

Aggregation Services Router (ASR) 11-26

AHT 25-5
ALI 15-3, 15-7, 15-29
alias normalization 14-75
ALI Formatting Tool (AFT) 15-29
all trunks busy 15-17
analog
  connection types 8-6
  endpoints 8-5
  gateways 5-2, 8-5
  interface modules 8-6
  standalone gateways 8-6
Analysis Manager 27-24
Analytics 27-12
anchoring calls in the enterprise 21-69
Android 8-37, 21-76, 21-90, 21-95
ANI 15-2, 15-7, 15-9, 15-14
annunciator 7-15
answer supervision 15-18
AnyConnect 21-109
AnyConnect Secure Mobility Client 8-38
AnyConnect VPN 21-103
AP 3-61, 3-63, 3-72, 8-33, 15-11
APNs 8-41, 21-99
Apple iOS 8-37, 21-99
Apple Push Notification service (APNs) 8-41, 21-99
application dialing rules 21-65
applications
  Attendant Console 18-42
  described 18-1
  Extension Mobility 18-7, 18-28
  for mobile users 21-1
  IP Manager Assistant 18-19
  IP Phone Services 18-2
  security 4-41
  Unified Communications Manager Assistant 18-19
  WebDialer 18-34
applications and services layer 17-1
application users 16-7
architecture
  applications and services layer 17-2
  call control and routing 12-2
  call processing 9-2
  Cisco Jabber 8-23, 20-7
  Cisco UC Integration for Microsoft Lync 8-27, 25-21
  Cisco UC Integration for Microsoft Office Communicator 25-21
  Cisco Unified Communications Manager Assistant 18-20, 18-22
  Cisco Unified Contact Center 22-2
  Cisco WebEx Connect 25-20
  collaboration system 2-1
  deployment models 10-4
  directories 16-7
  endpoints 8-2
  Enterprise Feature Access 21-67
  Extension Mobility 18-8
  IP Phone Services 18-2
  media resources 7-2
  Mobile Voice Access 21-67
  mobility clients and devices 21-77
  operations and serviceability layer 24-2
  presence 20-18
  Service Advertisement Framework (SAF) 10-59
  Single Number Reach 21-58
  trunks 6-2
  voice and video over WLAN 3-62
  WebDialer 18-34, 18-37
  wireless LAN 3-62
area code 14-80
ARP 3-72, 4-11
ASA 4-33, 4-39
ASR 11-26
Assistant Console 18-32
Assurance 27-4
Asynchronous Transfer Mode (ATM) 3-45, 10-15, 10-24
ATM 3-45, 10-15, 10-24
Attendant Console (AC) 18-42, 25-28
audio conferencing 11-4

audio on computer 8-25
audio sources 7-40
authentication
  database 3-64
  mechanisms 16-42
  of phones 4-29, 8-34
  of users 16-10, 16-22
  Security Assertion Markup Language (SAML) 16-37
authentication and encryption 4-29
authorization code grant flow 16-50
authorization framework 16-45
authorization grants 16-49
auto-detection 9-36
auto-generated directory numbers 16-17
automated alternate routing (AAR)
  dial plan considerations 14-70, 14-79
  for video calls 5-34
  for Voice over PSTN 10-22
  with Cisco Unity 19-7
automated attendant (AA) 19-22
automatic line creation 16-17
Automatic Location Identification (ALI) 15-7, 15-29
Automatic Location Identifier (ALI) 15-3
Automatic Number Identification (ANI) 15-2, 15-7, 15-9, 15-14
average hold time (AHT) 25-5
AXL 27-7

B

BackboneFast 3-6
bandwidth
  advanced formulas 3-59
  best-effort 3-36
  call control traffic 3-57, 3-58, 3-61
  consumption 3-52, 3-54, 3-55
  for Cisco Unity 19-32
  for conferencing 11-33
  for contact center 22-18
  for shared line appearances 3-59
  for video calls 13-66
  for WebEx 11-33
  general rule 10-44
  guaranteed 3-35
  management of 13-1
  provisioning 3-19, 3-35, 3-52
  requirements for call admission control 13-45
  voice class requirements 3-49
Basic Directory Integration (BDI) 8-32, 8-40
BDI 8-32, 8-40
BE4000 9-2, 9-26, 25-49
BE6000 9-2, 9-23, 25-49
BE7000 9-2, 9-23, 25-49
beacons 3-73
Bearer Capabilities Information Element (bearer-caps) 5-14
bearer-caps command 5-14
bearer traffic 3-53
best-effort bandwidth 3-36
Best Effort Early Offer 6-22, 6-24, 7-10
best practices for
  centralized call processing 10-16
  Cisco Unified Communications Manager Express (Unified CME) 9-38
  Cisco Unity 19-32
  Cisco Unity Connection 19-32
  Cisco Unity Express (CUE) 19-45
  distributed call processing 10-25
  LDAP synchronization 16-19
  music on hold 7-39
  single-site deployment 10-12
  voice messaging 19-32
  WAN design 3-34
BFD 11-31
BGP 11-31
BHCA 10-49, 25-5, 25-22, 25-50
BHCC 25-5
BIB 11-5, 23-6

Bidirectional Forwarding Detection (BFD) 11-31
bill-to number (BTN) 15-7
blade servers 10-56
BLF 20-16
blocking factor 25-5
blocking numbers 21-65
Bluetooth 3-70, 8-13, 8-20, 8-35, 8-40, 21-72
Border Gateway Protocol (BGP) 11-31
BPDU 3-6
branch office router 7-44
bridge protocol data unit (BPDU) 3-6
Bring Your Own Device (BYOD) Infrastructure 21-88
broadcast messages 18-47
B-Series Blade Server 10-56, 10-58
BTN 15-7
Built-in Bridge (BIB) 11-5, 23-6
bump in the wire 4-36
bursting 3-51
bursty traffic 25-6
Business Edition 9-2, 9-22, 9-23, 9-24, 9-26, 21-74, 25-49, 25-52
business-to-business communications 10-37
busy hour 25-5
busy hour call attempts (BHCA) 10-49, 25-5, 25-22, 25-50
busy hour call completions (BHCC) 25-5
busy lamp field (BLF) 20-16
BYOD 21-88

C

CAC (*see* call admission control)
calendar integration for presence 20-51
call admission control
  bandwidth management 13-1
  bandwidth requirements 13-45
  components 13-40
  described 13-1
  design considerations 13-73
  dual data center 13-74
  effective path 13-41
  elements 13-40
  enhanced locations 13-40
  example bandwidth deductions per call 13-62
  for contact center 22-19
  for music on hold 7-41
  for Session Management Edition (SME) 13-82
  for TelePresence 13-59, 13-78
  for video 13-66, 13-78
  links 13-41, 13-42
  locations 13-80
  migration to Enhanced Locations CAC 13-71
  moving devices to a new location 15-19, 21-15
  MPLS cloud 13-75
  paths 13-41
  regions 13-46, 13-47
  replication network 13-52
  SIP trunks 13-60
  topologies 13-73
  weights 13-41
call anchoring 21-69
callback
  for emergency services 15-15, 15-22
  from the PSAP 15-15, 15-22
call center 22-1
Call Control Discovery (CCD) 10-59
call control traffic 3-57, 3-61
call detail record (CDR) 10-46, 25-15, 27-8
caller ID matching 21-65, 21-66, 21-69
caller ID transformations 21-72
call flows
  multicast music on hold 7-23, 7-26
  music on hold 7-23, 7-26
  unicast music on hold 7-25, 7-28
Call Forward Unregistered (CFUR) 14-71
call hand-in 21-85
call handoff 21-85, 21-93
call hand-out 21-85, 21-93
calling line ID (CLID) 14-28
calling party number (CPN)

in 911 calls 15-7
localization 14-63
calling privileges 14-41
calling restrictions 14-41
calling search space 20-17
calling search spaces 14-41, 14-43, 20-17, 21-68
call management record (CMR) 10-46, 25-15, 27-8
call processing
agents 10-25
architecture 9-2
capacity planning 9-23
centralized 10-12, 19-6, 19-11, 22-12, 27-27
design considerations 9-26
distributed 10-23, 22-14, 27-28
guidelines 9-1
hardware platforms 9-4
high availability 9-13
redundancy 5-3, 9-14
subscriber server 9-6
Call Processing Language (CPL) 5-24
call routing
architectural layer 12-1
for emergency calls 15-26
inbound 21-80
outbound 21-81
calls
911 15-1
classification of 14-28
dual control 10-40
emergency 14-70, 15-1
forwarding 14-44
history 20-16
hold 7-19
inbound 5-32
monitoring 23-1
music on hold 7-17
outbound 5-33
pickup at desk phone 21-50
pickup at remote destination phone 21-51
point-to-point 21-99
preservation of 5-9
privileges 14-41
queuing 22-2
recording 23-1
routing 5-32, 5-33, 14-22, 15-26
signaling 5-14
Call Service Aware 21-41
Call Service Connect 21-41
calls per second (cps) 25-5
CAM 4-7
CAMA 15-9
campus
access switch 3-3
deployment model 10-10, 27-26
infrastructure requirements 3-1
capacity planning
applications and serviceability layer 17-4
Attendant Console 18-47, 25-28
Business Edition 9-23, 9-24, 25-49
by product 25-13
call processing 9-23
call recording and monitoring 23-10
call routing 12-3
call traffic 25-22
Cisco IM and Presence 25-33
Cisco mobility clients and devices 21-110
Cisco Prime Collaboration 25-48
Cisco Prime Collaboration Analytics 25-49
Cisco Prime Collaboration Assurance 25-48
Cisco UC Integration for Microsoft Lync 25-21
Cisco Unified Communications Manager Express (Unified CME) 9-26, 25-49
Cisco WebEx Connect 25-20
Cisco WebEx Messenger service 20-68
clusters 25-14
codecs 25-40
collaboration clients and applications 25-17
collaboration system 2-4

conferencing 25-44
contact center 22-21
CTI applications 9-32, 25-23
deployment models 10-6
design and deployment considerations 25-1
dial plan 25-23
emergency services 25-36
endpoints 8-44, 25-16
Extension Mobility 18-17, 25-26
factors to consider 25-9
gateways 25-38
instant messaging storage requirements 20-49
IP Phone Services 18-6
LDAP directory integration 25-31
locations 25-14
media resources 7-30, 25-28
megacluster 25-32
music on hold (MoH) 7-31, 7-33, 25-30
operations and serviceability layer 24-3
performance overload 25-40
performance tuning 25-41
phones 8-44
presence 25-33
regions 25-14
servers 25-14
sizing tools 25-10
tools 9-23, 25-10
Unified CM 25-13
Unified CM Assistant 18-26, 25-27
Unified CM servers 9-23
Unified MeetingPlace 25-44, 25-45
Unified Mobility 21-74, 25-21
videoconferencing 25-45
voice activity detection (VAD) 25-40
voice messaging 25-42
WebDialer 18-40, 25-27
WebEx 11-33
wireless networks 3-68, 8-34
XMPP clients 25-21
CAPWAP 3-63
CAR 10-46
CA-signed certificates 4-17
CCA 3-73, 11-31
CCD 10-59
CDI 8-32, 8-40, 21-92
CDP 4-5
CDR 10-46, 25-15, 27-8
CDR Analysis and Reporting (CAR) database 10-46
Centralized Automatic Message Accounting (CAMA) 15-9
centralized call processing
  centralized messaging 19-6
  deployment model 10-12, 22-12, 27-27
  distributed messaging 19-11
  migration to 26-5
  Voice over the PSTN 10-22
centralized IM and Presence deployment 20-32
centralized messaging 19-4, 19-6, 19-14, 19-21
CER 14-70, 15-10, 15-19
certificate management 4-14
Certificate Trust List (CTL) 4-23
CFUR 14-71
channels for wireless devices 3-69
chat rooms 20-41
CIR 3-51
Cisco AnyConnect VPN 21-103
Cisco Business Edition 9-2, 9-22, 9-23, 9-26, 21-74, 25-49, 25-52
Cisco Directory Integration (CDI) 8-32, 8-40, 21-92
Cisco Discovery Protocol (CDP) 4-5
Cisco Emergency Responder (CER) 14-70, 15-10, 15-19
Cisco EnergyWise Technology 3-13
Cisco Expressway 21-30, 25-37
Cisco IM and Presence 20-18, 25-33
Cisco IOS software MTP 7-14
Cisco IP Voice Media Streaming Application 7-15, 25-28
Cisco Jabber 8-23, 20-7, 21-90, 21-95
Cisco LEAP 8-34
Cisco Meeting Server 11-7

Cisco Mobile 21-90, 21-95
Cisco Mobile iPhone 21-95
Cisco Network Analysis Module (NAM) 27-9
Cisco Option Package (COP) 26-9
Cisco Paging Server 18-47
Cisco Prime 27-1
Cisco Prime Collaboration 25-48
Cisco Prime Collaboration Analytics 25-49
Cisco Prime Collaboration Assurance 25-48
Cisco Prime Unified Provisioning Manager (Unified PM) 27-13
Cisco Prime Unified Service Monitor (Unified SM) 27-8
Cisco Proprietary RTP 7-8
Cisco Spark 8-27, 8-37
Cisco Spark Room Series 8-17
Cisco UC Integration for Microsoft Lync 8-27, 25-21
Cisco UC Integration for Microsoft Office Communicator 25-21
Cisco Unified Analysis Manager 27-24
Cisco Unified Border Element 4-40
Cisco Unified Communications Management Suite 27-1
Cisco Unified Communications Manager Express (Unified CME)
  capacity planning 9-26, 25-49
  design considerations 9-28
  distributed call processing 10-25
  interoperability with Unified CM 9-36
Cisco Unified Communications Manager Real-Time Monitoring Tool (RTMT) 27-24
Cisco Unified Computing System (UCS) Platform 10-55
Cisco Unified Contact Center 22-1
Cisco Unified Contact Center Enterprise (Unified CCE) 22-3
Cisco Unified Contact Center Express (Unified CCX) 22-6
Cisco Unified Contact Center Management Portal (Unified CCMP) 22-8
Cisco Unified Customer Voice Portal (Unified CVP) 22-4
Cisco Unified E-Mail Interaction Manager (Unified EIM) 22-9
Cisco Unified Intelligence Center (Unified IC) 22-9
Cisco Unified MeetingPlace 25-44, 25-45
Cisco Unified Mobility 21-1, 21-47, 21-107, 25-21, 25-52
Cisco Unified Reporting 27-25
Cisco Unified SRST Manager 10-21
Cisco Unified Survivable Remote Site Telephony (SRST) Manager 10-21
Cisco Unified Web Interaction Manager (Unified WIM) 22-9
Cisco Unity 19-1, 19-6, 19-19
Cisco Unity Connection 19-6, 19-17, 19-34
Cisco Unity Express (CUE) 19-22
Cisco Unity Personal Assistant 19-4
Cisco Unity Telephony Integration Manager (UTIM) 19-40, 19-42
Cisco Voice Transmission Quality (CVTQ) 27-8
Cisco WebEx Connect 25-20
Cisco WebEx Meeting Center Video Conferencing 11-34
Cisco WebEx Meetings Server 11-41
classification of
  calls 14-28
  traffic 3-4, 3-16, 3-75
Class of Service (CoS) 3-4
clear channel assessment (CCA) 3-73
CLEC 15-6
CLID 14-28
Client Matter Code (CMC) 14-29
clients
  mobility clients and devices 21-76
clipping 10-16
cloud architecture 11-26
cloud-based deployment model 20-12
Cloud Connected Audio (CCA) 11-31
cloud services 21-34
clustering over the WAN
  Cisco Unity 19-14, 19-16
  CTI applications 9-31
  described 10-43
  failover with Cisco Unity 19-18
  for contact center 22-15, 27-29
  local failover 10-47

music on hold 7-47
presence 20-29
remote failover 10-54
troubleshooting 10-47
WAN considerations 10-44
with Cisco Unity 19-19
clusters
design guidelines 9-5
Emergency Responder (ER) 15-13, 15-26
for presence servers 20-19
for Unified CM 9-5
guidelines for 9-12
home 18-14
home cluster 18-18
maximum capacity 25-14
redundancy 9-16
server nodes 9-6
services 9-5
visiting 18-14
CMC 14-29
CMR 10-46, 11-34, 11-49, 25-15, 27-8
CMR Hybrid
personal meeting room 11-49
codecs
capacity planning 25-40
complexity modes 7-4
flex mode 7-4
for music on hold 7-39
low bit-rate (LBR) 7-37
collaboration
clients 20-5
clients and applications 25-17
conferencing 25-44
contact management 8-26
Jabber desktop clients 8-23, 20-7
LDAP directory integration 8-26, 20-9
third-party XMPP clients and applications 25-21
Collaboration Cloud 11-26
Collaboration Meeting Room (CMR) 11-49
Collaboration Meeting Rooms (CMR) 11-34
Collaboration Sizing Tool 9-23, 25-10
collaboration system components and architecture 2-1
collaborative conferencing 25-44
co-located DHCP server 3-26
COM 16-4
combined deployment models for messaging 19-13
Committed Information Rate (CIR) 3-51
common locations 13-53
Communicator 8-22
competitive local exchange carrier (CLEC) 15-6
complexity modes for codecs 7-4
complexity of the database 25-14
Component Object Model (COM) 16-4
components of
Device Mobility 21-16
messaging system 19-2
presence 20-3
compressed Real-Time Transport Protocol (cRTP) 3-46, 3-48
Computer Telephony Integration (CTI) 9-7, 9-20, 9-28, 19-22, 25-23
Conference Now 11-5
conferencing
collaborative 25-44
conference bridges 7-14
described 11-1
hardware 9-37
rich media 11-1
security 4-40
traffic 25-8
configuration examples for
lobby phone security 4-43
Unified CME 9-36
configuration for mobile client users
simplified method 21-87
conformance with Section 508 8-5
connectivity options for the WAN 10-15, 10-24
console

for attendants 18-42
for Unified CM Assistant assistant 18-32
contact center
described 22-1
gateway sizing 25-39
traffic patterns 25-7
contact lists 20-59
contact management 8-26, 20-59
Contact Sharing 22-10
contact sources 8-32, 8-40
content-addressable memory (CAM) 4-7
Context Service 22-10
Control and Provisioning of Wireless Access Points (CAPWAP) 3-63
control signaling 3-57, 3-61
COP 26-9
Core Layer 3-11
co-resident
DHCP 3-27
MoH 7-31
core switch 3-3
CoS 3-4
CPL 5-24
CPN 15-7
cps 25-5
CPU usage 25-4
cRTP 3-46, 3-48
C-Series Rack-Mount Server 10-58
CTI 9-7, 9-20, 9-28, 19-22, 25-23
CTI Manager 9-5, 9-7, 9-20
CTI-QBE 19-22
CTI Remote Device 9-28
CTI route points 7-13
CTL 4-23
CUE 19-22
customer care using video 22-22
cutover 26-1
CVTQ 27-8

D

DAI 4-10, 4-11
database
complexity 25-14
replication 9-9
synchronization with Unified CM 16-31
data centers
security 4-37
server farm 3-12
Delayed Offer 6-18, 7-9
delay of packets 10-44, 10-46
Delivery Traffic Indicator Message (DTIM) 3-71
Demilitarized Zone (DMZ) 4-44
deployment models
campus 10-10, 27-26
clustering over the WAN 7-47, 10-43, 19-19, 20-29, 22-15, 27-29
combined for messaging 19-13
described 10-1
DHCP 3-26
federation 20-36
for Cisco Jabber 20-10
for Cisco Unity 19-3
for Cisco Unity Express 19-22
for contact center 22-12
for network management 27-26
for presence 20-26
for presence servers 20-22
for Unified CME 9-38
media resources 7-36
messaging and call processing combinations 19-5
multisite with centralized call processing 7-37, 7-43, 10-12, 22-12, 27-27
multisite with distributed call processing 7-38, 7-47, 10-23, 22-14, 27-28
music on hold 7-43
Service Advertisement Framework (SAF) 10-59
Session Management Edition 10-26
single cluster 20-26

single site 7-36, 7-43, 10-10, 22-12, 27-26
site-based 10-6
Unified Computing System (UCS) 10-55
virtualized servers 10-55, 10-59
voice over the PSTN 10-22
design criteria 10-6
designing for performance 25-9
deskphone control mode (using deskphone for audio) 8-25
deskphone for audio 8-25
desk phone pickup 21-50
desk phones 8-8
destination of a call 14-80
device location discovery 15-10
device mobility
dial plan 21-21
feature components and operation 21-16
Group 21-16
Info 21-16
operation flowchart 21-20
operation of 21-20
parameter settings 21-18
Physical Location 21-16
settings 21-19
Device Mobility Group 21-19
devices
mobility 8-36, 15-19, 21-15
pools 10-48, 10-54
route group 14-30
Device Security Profile 18-14
DFS 3-69
DHCP
binding information 4-10
deployment options 3-26
described 3-24
lease times 3-25
Option 150 3-25
servers 3-27
Snooping 4-8, 4-10
starvation attack 4-10
dial plan
+ dialing 14-57
911 calls 15-1
application dialing rules 21-65
architecture 14-3
Call Forward Unregistered (CFUR) 14-71
calling party settings 14-59
calling privileges 14-41
call routing 14-22
capacity planning 25-23
design considerations 21-21
device mobility 21-21
elements 14-13
emergency call string 15-16
Extension Mobility 14-84
for Device Mobility 21-21
for mobility 21-82
for software-based endpoints 8-31
for Unified CM Assistant 18-29
functions 14-1
fundamentals 14-3
globalized numbers 14-56, 14-62
international calls 14-27
localized call egress 14-63
localized call ingress 14-61
local route group 14-57
protection 5-24
shared line appearance 15-22
Tail End Hop Off (TEHO) 14-71
transformations 14-58
Unified Mobility 21-68
variable length on-net dialing 21-23
Video Communication Server (VCS) 14-53
dial rules 14-16, 14-18, 14-20, 21-65
dial via office (DVO) 21-86, 21-96
dial via office forward (DVO-F) 21-99
dial via office reverse (DVO-R) 21-97
DID 15-7

Differentiated Services Code Point (DSCP) 3-4, 3-47, 3-75, 13-81
digital gateways 5-3
digital networking 19-29
digital signal processor (*see* DSP resources)
digit manipulation 5-33, 14-24, 14-28
digit prefixing 21-66
Direct Inward Dial (DID) 15-7
directories
  access 16-4, 16-6, 21-92
  architecture 16-7
  authentication of users 16-10, 16-22
  filtering 16-28
  for Unified CM Assistant 18-33
  high availability 16-31
  integration with IP telephony system 16-1, 16-3, 25-31
  integration with Unified CM 16-7
  LDAP 16-1, 25-31
  schema 16-1
  search base 16-13
  searches 8-27
  security 16-19
  sn attribute 16-10
  synchronization 16-10, 16-28
  URI dialing 14-23, 14-50
  UserID 16-10
directory numbers, auto-generated 16-17
directory URI 14-49
distortion 3-70
distributed call processing 10-23, 10-25, 22-14, 27-28
distributed messaging 19-4, 19-11, 19-16
Distribution Layer 3-9
DMVPN 3-35
DMZ 4-44
DNS 3-23
Domain Name System (DNS) 3-23
DSCP 3-4, 3-47, 3-75, 13-81
DSP resources
  described 7-4
  PVDM 7-30
DTIM 3-71
DTMF
  conversion of 7-7
  gateway capabilities 5-3
  methods supported by endpoints 7-7
  on H.323 gateways 7-13
  on SIP gateways 7-12
  Relay 5-5, 7-13
DTPC 3-72
dual call control 10-40
dual data center 13-74
dual-mode
  clients 21-90, 21-95
  phones and clients 21-76
dual tone multifrequency (DTMF) 5-3, 5-5, 7-7
duplex media 7-30
duplex unicast MoH 7-30
DVO 21-86, 21-96
DVO-F 21-99
DVO-R 21-97
DX6 Series video endpoins 8-15
DX Series video endpoins 8-10
dynamic ANI interface 15-14
Dynamic ARP Inspection (DAI) 4-10, 4-11
Dynamic Frequency Selection (DFS) 3-69
Dynamic Host Configuration Protocol (DHCP) 3-24, 4-8, 4-10
dynamic memory 25-4
Dynamic Multipoint VPN (DMVPN) 3-35
Dynamic Transmit Power Control (DTPC) 3-72

E

E.164 15-7, 15-14, 19-37
E911 15-1, 15-6
Early Offer 6-19, 7-9
ECC variables 22-8
ECDSA 4-16

EDI 8-32, 8-40
effective path 13-41
efficiency of links 3-48
ELCAC 13-40, 13-85
elements of a dial plan 14-13
ELIN 15-13, 15-14
Elliptical Curve Digital Signature Algorithm (ECDSA) 4-16
EMCC 18-9, 18-18, 25-26
emergency call routing 15-27
emergency calls 14-70, 15-1
emergency call string 15-16
emergency location identification number (ELIN) 15-13, 15-14
Emergency Responder 14-70, 14-71, 15-10, 15-19
emergency response location (ERL) 15-13, 15-14, 15-19
emergency services 15-1, 21-83, 25-36
eMWI 19-38
encryption
  for phones 4-29
  for security 4-19, 4-29
  for signaling 3-58, 3-59
  for wireless endpoints 8-34
endpoints
  analog gateways 8-5
  architecture 8-2
  capacity planning 8-44, 25-16
  design considerations 8-44
  directory access 16-4
  high availability 8-43
  immersive video 8-18
  mobile 8-37
  multipurpose video 8-16
  off premises 15-21
  personal video 8-15
  Section 508 conformance 8-5
  security 4-25
  software-based 8-22
  supplementary services 7-12
  telepresence 4-28, 8-16, 8-17, 8-18
  types of 8-1
  video 8-14, 15-20
  wireless 3-65, 8-33
end users 16-7, 20-3
Energy conservation 3-13
EnergyWise Technology 3-13
Enhanced Directory Integration (EDI) 8-32, 8-40
Enhanced Location CAC 13-40, 13-85
Enhanced Message Waiting Indicator (eMWI) 19-38
Enhanced SRST 8-13, 8-30, 8-36, 8-40, 8-42
Enhanced SRST (E-SRST) 8-19, 10-16
Enhanced SRST (E-SRST) 10-19
Enhanced Survivable Remote Site Telephony (E-SRST) 10-16
enterprise caller ID 21-83
Enterprise Feature Access 21-46, 21-52, 21-63, 21-65
enterprise groups 16-19
equations for calculating
  bandwidth 3-57, 3-59
  Business Edition device capacities 25-50
  CPU usage 25-4
  CTI resource requirements 25-24
  memory usage 25-4
  music on hold server capacity 7-32
ERL 15-13, 15-14, 15-19
Erlang 25-6
Erlang blocking factor 25-6
error rate 10-47
E-SRST 8-19, 10-16, 10-19
ESXi Hypervisor. 26-10
eTokens 4-22
ettercap virus 4-11
Exchange Web Services Calendar 20-53
Expressway 4-41, 13-85, 21-30, 21-101, 21-103, 25-37
EX Series video endpoints 8-16
Extend and Connect 8-32
Extended Call Context (ECC) 22-8
Extensible Authentication Protocol (EAP) 8-34

extensible messaging 20-57
Extension Mobility (EM)
capacity planning 25-26
described 18-7
dial plan 14-84
interactions with Unified CM Assistant 18-28
Extension Mobility Cross Cluster (EMCC) 18-9, 18-18, 25-26
EMCC 13-73
external MoH source 7-22

F

FAC 14-29
factors that affect sizing 25-9
failover
Cisco Unity 19-17, 19-18
clustering over the WAN 10-47, 10-54
scenarios 18-5
fallback mode 7-46
Fast Start 7-12
fax
gateway support for 5-3, 5-37
interface modules 8-6
FCoE 10-56, 10-57
Feature Group Template 16-17
federated deployment 20-36
federation between domains 20-36
Fibre Channel over Ethernet (FCoE) 10-56, 10-57
filtering for directory synchronization and authentication 16-28
filter strings for LDAP directories 16-31
Finesse 22-7
firewalls
access control lists 20-68
around gateways 4-39
bump in the road 4-36
centralized deployment 4-44
described 4-33
routed mode 4-35
stealth mode 4-36
transparent mode 4-36
Firewall Services Module (FWSM) 4-33, 4-39
firmware upgrades for Cisco IP Phones 8-11
flash used for music on hold 7-44
flat addressing 21-23
flex mode for codecs 7-4
Forced Authorization Codes (FAC) 14-29
Foreign Exchange Office (FXO) 15-9
forwarding calls 14-44
Frame Relay 3-45, 10-15, 10-24
FWSM 4-33, 4-39
FXO 15-9

G

gain settings 5-32
GARP 4-11
gatekeeper
call admission control 10-25
Gateway Load Balancing Protocol (GLBP) 3-10
gateways
911 services 15-17
additional documentation 25-42
all trunks busy 15-17
analog 5-2, 8-5
automated alternative routing 5-34
blocking 15-17
call recording 23-7
capabilities 5-14
capacity planning 25-38
Cisco Unified Videoconferencing 3500 Series Video Gateways 5-11
configuration in Unified CM 5-13
contact center sizing 25-39
core feature requirements 5-5
digital 5-3
digit manipulation 5-33

firewalls 4-39
for local failover 10-53
for video telephony 5-11
placement 15-17
protocols 5-3
redundancy 5-9
security 4-38
selection of 5-3
service prefixes 5-34
SIP 5-6, 5-11
standalone 8-6
types of 5-2
voice applications 5-1, 8-5
VoiceXML 21-60, 21-61
GDPR 14-11, 14-47, 14-72
general security 4-2
GeoDNS 5-26
geographical diversity 10-9
geolocations 14-92
GLBP 3-10
GLO 26-7
Global Dial Plan Replication (GDPR) 14-11, 14-47, 14-72
globalized dial plan 14-56, 14-62
Global Licensing Operations (GLO) 26-7
Global Site Backup (GSB) 11-26, 11-30
glossary 1-1
grant flows 16-49
Gratuitous Address Resolution Protocol (GARP) 4-11
ground start 8-6
groups for
call routing 14-30
Emergency Responder (ER) 15-22, 15-24
gateways 25-38
media resources 7-1
Unified CM redundancy 9-14
GSB 11-26, 11-30
guaranteed bandwidth 3-35

H

H.245 Alphanumeric 7-8
H.245 Signal 7-8
H.323
call hairpinning 9-36
call preservation enhancements 5-9
Fast Start 7-12
gateways 5-3
supplementary services 7-12
trunks 6-3
hairpinning 9-36, 21-61
hand-in of a call 21-85
handoff of calls 21-85, 21-93
hand-out of a call 21-85, 21-93
hardware
media resource capacities 7-30
MTP resources 7-15
music on hold 7-31
types of platforms 9-4
hardware USB eTokens 4-22
headers for voice packets 3-53
high availability
applications and serviceability layer 17-3
Attendant Console 18-45
Business Edition 9-22
call processing 9-13
call routing 12-3
Cisco mobility clients and devices 21-109
collaboration system 2-4
contact center 22-17
CTI 9-32
deployment models 10-5
directories 16-31
endpoints 8-43
Enterprise Feature Access 21-68
Extension Mobility 18-15
hardware platforms 9-13
IP Phone Services 18-5

media resources 7-34, 7-35
Mobile Voice Access 21-68
music on hold 7-36
network connectivity 9-13
network services 3-4
operations and serviceability layer 24-3
phones 8-43
presence 20-21
requirements 10-7
Single Number Reach 21-58
Survivable Remote Site Telephony (SRST) 9-16
transcoders 7-36
Unified CM 9-14
Unified CM Assistant 18-24
Unified Computing System (UCS) 9-21
voice services 10-16
WebDialer 18-39
WebEx 11-30
wireless LAN 3-66
history of calls 20-16
hold 7-17, 7-19
holdee 7-18
holder 7-18
home cluster 18-14, 18-18
Hot Standby Router Protocol (HSRP) 3-10, 10-25
HSRP 3-10, 10-25
HTTPS 19-30
hub-and-spoke topology 3-3, 3-34
hybrid deployment model 20-13
hybrid services 21-34
hypervisor 3-20, 10-55

I

I/O modules 10-57
IButton 14-21
ICCS 9-9, 10-45, 10-49
ICMP 5-11
identity management 16-1, 16-33
Identity Provider (IdP) 16-33
IdP 16-33
IDS 4-39, 10-45
IM and Presence 20-1, 25-33
immediate start 8-6
immersive video endpoints 8-18
impairments without QoS 3-19
implicit grant flow 16-49
IM push notifications 21-99
inbound calls 5-32
InformaCast 18-47
Informix Dynamic Server (IDS) 10-45
infrastructure (*see* network infrastructure)
Initial Trust List (ITL) 4-23
inline power 3-12
instant messaging 20-1, 20-41, 20-49
Intelligent Proximity 8-13, 8-20, 8-40, 21-72, 21-107
Intelligent Session Control 21-70
interactive voice response (IVR) 10-12
interface modules 8-6
interface types for 911 calls 15-7
interference to wireless communications 3-70
international calls 14-27
Internet Control Message Protocol (ICMP) 5-11
interoperability 8-20, 9-36, 9-40, 13-78
inter-VLAN routing 8-19, 8-30
Intra-Cluster Communication Signaling (ICCS) 9-9, 10-45, 10-49
introduction 1-1
Intrusion Detection System (IDS) 4-39
IOS software MTP 7-14
IP/VC 3500 Series Video Gateways 5-11
IP addresses and security 4-4
IP Communicator 8-22
iPhone 8-37, 21-76, 21-90, 21-95
IPMA 18-19
IP Manager Assistant (IPMA) 18-19
IP phones 8-8
IP Phone Services 18-2, 25-25

IP Precedence 3-4, 3-47
IPSec 10-15, 10-24
IP Security Protocol (IPSec) 10-15, 10-24
IPv6
security 4-5
with Cisco Unified Provisioning Manager (Unified PM) 27-17
with Cisco Unity Connection 19-43
IPVMS 25-28
IP VOICE feature set 9-36
IP Voice Media Streaming Application 7-3, 7-14, 7-15, 25-28
ISDN 10-16, 10-17
ISDN Link 5-3
ITL 4-23
IVR 10-12
IX5000 Series immersive video system 8-18

J

Jabber
call handoff 21-93
clients 25-18
deployment models 20-10
Desktop Client Cache 8-27
desktop clients 8-23, 20-7, 25-17
desktop video 8-15
dial via office (DVO) 21-96
for Android and Apple iOS 8-37, 21-90
for mobile devices 21-76
interactions with Cisco Unified Mobility 21-107
WLAN considerations 21-95
Jabber Identifier (JID) 20-3
Jabber Service Discovery 21-91
JID 20-3
jitter 10-44
JTAPI 9-20

K

Key Press Markup Language (KPML) 7-8, 14-16, 14-18
KPML 7-8, 14-16, 14-18

L

LAN infrastructure 3-4
Layer 2 3-4, 10-25
Layer 3 3-4
layers of security 4-3
LBM 13-41, 13-48
LBM Hub 13-41, 13-52
LBR 7-37
LCR 5-36
LDAP 8-26, 8-27, 9-9, 16-1, 16-32, 20-9, 25-31
LDN 15-7
LEAP 8-34
leased lines 3-45, 10-15, 10-24
lease times for DHCP 3-25
least-cost routing (LCR) 5-36
LEC 15-2, 15-5, 15-17
LFI 3-46, 3-48, 3-49
Lightweight Access Point Protocol (LWAPP) 3-63
Lightweight Directory Access Protocol (LDAP) 9-9, 16-1, 16-32, 25-31
Lightweight Directory Services 16-22
Limit Client Power setting on access points 3-72
line appearances 3-59
line speed mismatch 3-51
link efficiency 3-48
link fragmentation and interleaving (LFI) 3-46, 3-48, 3-49
links for call admission control 13-41, 13-42
listed directory number (LDN) 15-7
Live Communications Server 2005 20-62
LLQ 3-46, 3-47
LMHOSTS file 3-23
load balancing 3-31, 9-19
lobby phone security 4-43

Local Exchange Carrier (LEC) 15-2, 15-5, 15-17
local failover deployment model 10-47
localization of calling party number 14-63
localized call egress 14-63
localized call ingress 14-61
local route group 14-31, 14-57
Location and Link Management Cluster 13-56
location discovery for emergency calls 15-10
locations
  common 13-53
  defined 13-41
  enhanced 13-40
  for video endpoints 13-80
  maxiumum number 25-14
  shadow location 13-55
  shared 13-53
Locations Bandwidth Manager (LBM) 13-41, 13-48
Locations Bandwidth Manager Hub 13-41, 13-52
logical partitioning 14-60, 14-92
loop start 8-6
low bit-rate (LBR) codecs 7-37
low-latency queuing (LLQ) 3-46, 3-47
LWAPP 3-63
Lync 8-27

M

MAC address 4-7
managed file transfer (MFT) 20-44
manipulation of digits 14-24
Master Street Address Guide (MSAG) 15-3
maximum simultaneous calls 25-5
MDM 20-6
Mean Opinion Score (MOS) 27-8
Media Gateway Control Protocol (MGCP) 5-3
media resource group (MRG) 7-34
media resource group list (MRGL) 7-34
Media Resource Manager (MRM) 7-2
media resources
  architecture 7-2
  capacity planning 7-30, 25-28
  deployment models 7-36
  described 7-1
  design guidelines 7-34
  for local failover 10-53
  hardware and software capacities 7-30
  high availability 7-34, 7-35
  PVDM 7-30
  security 4-38
  server 9-7
  voice quality 7-39
Media Routing Domain (MRD) 22-7
Media Streaming Application 7-3, 7-14, 7-15, 25-28
media termination point (MTP)
  conference bridges 7-14
  described 7-7
  types 7-14
  with SIP trunk 6-6
media transparency 6-24
meeting room, personal 11-34, 11-49
Meeting Server 11-7
megacluster 9-25, 10-4, 25-32
memory usage 25-4
Message Waiting Indicator (MWI) 19-22
messaging
  bandwidth management 19-32
  centralized 19-4, 19-6, 19-14, 19-21
  Cisco Unity 19-1
  combined deployment models 19-13
  deployment models 19-3
  distributed 19-4, 19-11, 19-16
  failover 19-17, 19-18
  redundancy 19-17
  system components 19-2
MFT 20-44
MGCP 5-3
Microsoft Active Directory (AD) 16-10, 16-15, 16-20, 16-26

Microsoft Active Directory Application Mode (ADAM) 16-12, 16-31
Microsoft Communications Server 20-62
Microsoft Lync 8-27, 25-21
Microsoft Office Communicator 20-62
Microsoft ViewMail for Outlook (VMO) 19-4
mid-call features 21-52, 21-84
migration
to Enhanced Locations CAC 13-71
to IP Telephony 26-1
to Unified CM 26-1
MISTP 3-4
mixed mode 4-22
MLP 3-46
MLPP 7-15
MLTS 15-2
mobile and remote access 21-101, 21-103
mobile and remote access (MRA) 16-52
Mobile Connect
described 21-46
mobile endpoints 8-37
Mobile Voice Access
access numbers 21-65
architecture 21-67
described 21-46, 21-59, 21-72
functionality 21-60
hairpinning 21-61
IVR VoiceXML gateway 21-60
number blocking 21-65
redundancy 21-68
Mobile Voice capabilities 8-13, 8-20, 8-40, 21-107
Mobility
applications 21-1
clients and devices 21-76
cloud services 21-34
described 21-1, 21-68
dial plan 21-82
emergency services 21-83
guidelines for deploying 21-73
hybrid services 21-34
integration with presence 20-55
softkey method of call hand-out 21-93
voicemail avoidance 21-55
modeling of computer systems 25-3
models for deployments (*see* deployment models)
modems, gateway support for 5-3, 5-37
MoH 7-17, 10-53, 25-30
monitoring calls 23-1
MOS 27-8
moves, adds, and changes 15-10
MPLS 3-33, 3-45, 10-15, 10-24
MPLS cloud 13-75
MRA 16-52
MRD 22-7
MRG 7-34
MRGL 7-34
MRM 7-2
MSAG 15-3
MTLS 4-20
MTP
conference bridges 7-14
described 7-7
hardware resources 7-15
software resources 7-14
types 7-14
with SIP trunk 6-6
multicast music on hold 7-17, 7-22, 7-23, 7-26, 7-39, 7-40, 7-44
multicast traffic on WLAN 3-71
multicast voice messages 18-47
multichannel support 22-9
multi-forest LDAP synchronization 16-22
Multilevel Precedence Preemption (MLPP) 7-15
multi-line telephone system (MLTS) 15-2
Multilink Point-to-Point Protocol (MLP) 3-46
multipath distortion 3-70
Multiple Device Messaging (MDM) 20-6
Multiple Instance Spanning Tree Protocol (MISTP) 3-4
multiple local route groups 14-34

multiple Unified CM servers 19-21
Multiprotocol Label Switching (MPLS) 3-33, 3-45, 10-15, 10-24
multipurpose video endpoints 8-16
multi-server certificates 4-18
multisite deployment model
with centralized call processing 7-37, 7-43, 10-12, 22-12, 27-27
with distributed call processing 7-38, 7-47, 10-23, 22-14, 27-28
music on hold (MoH) 7-17, 10-53, 25-30
Mutual TLS (MTLS) 4-20
MWI 19-22
MX Series video endpoints 8-16

N

NAM 27-9
Named Telephony Event (NTE) 5-6, 7-7
NAT 4-37
National Emergency Number Association (NENA) 15-13, 15-29
Native Emergency Call Routing 15-27
native interoperability for video 13-78
native transcoding with Cisco Unity 19-33
NENA 15-13, 15-29
Network Address Translation (NAT) 4-37
Network Analysis Module (NAM) 27-9
network hold 7-19
network infrastructure
access layer 3-4
core layer 3-11
distribution layer 3-9
high availability 3-4
LAN 3-4
network management 27-4
requirements 3-1
roles 3-3
routed access layer 3-7
security 4-4
voice over wireless LAN (WLAN) 21-78
WAN 3-33
wireless LAN 21-78
WLAN 3-61
network management 22-23, 27-1
network services 3-23
Network Time Protocol (NTP) 3-33
Network Transmission Loss Plan (NTLP) 5-32
Nexus 1000V Switch 3-20
non-fallback mode 7-44
normalization
of aliases 14-75
NPA 14-80
NTE 5-6, 7-7
NTLP 5-32
NTP 3-33
number blocking 21-65
Numbering Plan Area (NPA) 14-80
number transformations 14-58
numeric URI 14-49, 14-52

O

OAuth 2.0 8-32, 8-41, 16-45, 21-101
Office Communications Server 2007 20-62
off-premises endpoints 15-21
on-premises deployment model 20-11
OpenAM 20-4
open authentication 8-34
Open Shortest Path First (OSPF) 4-35
Open Virtualization Archives (OVA) 9-27
operations and serviceability layer 24-1
Option 150 3-24, 3-25
OSPF 4-35
outbound calls 5-33
OVA templates 9-27
overlap
of channels 3-69
receiving 14-28

sending 14-28
oversubscription of a link 3-51

P

packets
delay 10-44, 10-46
headers 3-53
jitter 10-44
loss of 10-44
Paging Server 18-47
paging systems 8-7
parallel cutover 26-3
parameters for Device Mobility 21-18
partial caller ID matching 21-66
partitions 14-41, 14-42, 14-60, 14-92
passive-interface command 3-11
paths for call admission control 13-41
PC port on IP phone 4-26
performance
call rate 9-1
designing for 25-9
modeling 25-3
of call processing servers 9-23
of Extension Mobility 18-17
of presence servers 20-26
of Unified CM Assistant 18-26
of WebDialer 18-40
overload on gateways 25-40
tuning of gateways 25-41
performance testing 25-2
persistent chat 20-31, 20-41, 20-49
Personally Identifiable Information (PII) 22-10
personal meeting room 11-34
personal video endpoints 8-15
phased migration 26-3
phone books 27-20
phones
3900 Series 8-10
7800 Series 8-8
7900 Series 8-8
8800 Series 8-9, 8-15
Attendant Console 18-42
authentication and encryption 4-29
call pickup at desk phone 21-50
capacity planning 8-44
design considerations 8-44
desktop IP models 8-8
dual-mode 21-76, 21-109
energy conservation 3-13
Extension Mobility 18-7
firmware upgrades 8-11
high availability 8-43
IP Phone Services 18-2
mid-call features 21-52
PC port 4-26
Power Save mode 3-14
Power Save Plus mode 3-13
remote destination call pickup 21-51
roaming 3-69
SCCP 14-15
secure mode 18-14
security 4-25, 4-43
services 18-2, 25-25
settings 4-28
SIP 8-43, 14-16, 14-18
software-based 8-22
Type-A 14-16
Type-B 14-18
Unified Communications Manager Assistant 18-19
user input 14-15, 14-16, 14-18
web access 4-27
WebDialer 18-34
wireless 8-33
Wireless IP Phone 7921G 8-33
Wireless IP Phone 7925G 8-33
Wireless IP Phone 7925G-EX 8-33
Wireless IP Phone 7926G 8-33

physical security 4-4
Piece of Data (POD) 22-10
PII 22-10
ping utility 10-46
PIX 4-33, 4-39
PKI 4-14
plain old telephone service (POTS) 15-9
platforms 9-4
POD 22-10
PoE 3-12, 8-12
point-to-point calling 21-99
policy
  for network security 4-2
  for presence 20-17
polling model 20-56
PortFast 3-6
ports
  access 4-7
  for integration of Cisco Unity with Unified CM 19-40, 19-42
  on the IP phone 4-26
  security 4-6
POTS 15-9
Power over Ethernet (PoE) 3-12, 8-12
Power Save mode 3-14
Power Save Plus mode 3-13
precedence settings for network traffic 3-4, 3-47
prefixes
  for access code 14-80
  service 5-34
presence
  calendar integration 20-51
  call history 20-16
  capacity planning 25-33
  clustering over the WAN 20-29
  clusters 20-19
  components 20-3
  contact lists 20-59
  deployment models 20-22, 20-26
  described 20-1, 20-2
  end user 20-3
  Exchange Web Services Calendar integration 20-53
  federation 20-36
  groups 20-17
  guidelines 20-18
  instant messaging storage requirements 20-49
  integration with third-party applications 20-62
  interactions between components 20-26
  message archiving and compliance 20-46
  Microsoft Communications Server 20-62
  migration 26-14
  mobility integration 20-55
  policy 20-17
  polling model 20-56
  presentity 20-2
  protocol interfaces 20-57
  real-time eventing model 20-55
  server guidelines 20-58
  server performance 20-26
  server redundancy 20-21
  servers 20-18
  server synchronization 20-19
  SIP 20-14
  speed dial 20-16
  state changes 20-60
  SUBSCRIBE calling search space 20-17
  synchronization of servers 20-19
  Third-Party Open API 20-55
  Unified CM 20-14
presentity 20-2
preservation of calls 5-9
PRI 15-7
primary extension 20-3
Primary Rate Interface (PRI) 15-7
Prime Collaboration 25-48, 27-2
Prime Collaboration Analytics 25-49, 27-12
Prime Collaboration Assurance 25-48
Prime Collaboration Deployment 26-3

Prime compliance 27-1
prioritization of traffic 3-47
private certificate authority 4-19
Private Internet Exchange (PIX) 4-33, 4-39
Private Switch ALI 15-4
privileges for making calls 14-41
progress_ind alert enable 8 command 15-18
propagation of database 9-9
protocols
ARP 3-72, 4-11
BFD 11-31
BGP 11-31
CAPWP 3-63
CDP 4-5
cRTP 3-46, 3-48
DHCP 3-24, 4-8, 4-10
GARP 4-11
GLBP 3-10
H.323 5-3, 6-3, 9-36
HSRP 3-10, 10-25
IPSec 10-15, 10-24
LDAP 9-9, 16-1, 25-31
LWAPP 3-63
MGCP 5-3
MISTP 3-4
MLP 3-46
NTP 3-33
RCP 4-11
RIP 4-35
routing 3-11
RSTP 3-4, 3-7
RSVP 3-34
RTP 10-25
SCCP 5-3, 7-8, 7-23, 14-15
SIMPLE 20-18
SIP 5-6, 5-11, 6-3, 6-5, 6-6, 7-16, 7-26, 8-43, 9-40, 10-25, 14-16, 14-18, 14-20, 20-14
SMTP 19-28
SNMP 15-10
SOAP 20-19
SRTP 3-53, 4-29
STP 3-6
TFTP 3-25, 3-28, 9-5, 9-20
TLS 4-29
UDP 10-25
VPIM 19-28
VRRP 3-9
provisioning servers 9-23
proxy
line mode with Unified CM Assistant 18-20
proxy TFTP 3-32
PSAP 15-2, 15-15, 15-22
PSTN
911 calls 15-2
access to remote sites 10-15, 10-24
destination number 14-80
traffic patterns 25-39
voice over the PSTN (VoPSTN) 10-22
public certificate authority 4-19
public key infrastructure (PKI) 4-14
public safety answering point (PSAP) 15-2, 15-15, 15-22
Public Switched Telephone Network (PSTN) 10-15, 10-24, 14-80, 15-2
publisher server 9-6, 10-45
push notifications 21-99
PVDM 7-30

Q

QBE 9-29, 19-22
QBSS 3-73, 3-77
QoS
for analog endpoints 8-7
for Cisco Unified Computing System (UCS) 3-20
for contact center 22-18
for desk phones 8-12
for LAN 3-14
for mobile clients and devices 21-80

for mobile endpoints 8-39
for music on hold 7-41
for security 4-31
for software-based endpoints 8-29
for Unified CM Assistant 18-32
for video 8-22, 13-78
for video endpoints 8-18
for WAN 3-33, 3-37
for wireless endpoints 8-36
for wireless LAN 3-74
QoS Basic Service Set (QBSS) 3-73, 3-77
Quality of Service (QoS)
for analog endpoints 8-7
for Cisco Unified Computing System (UCS) 3-20
for contact center 22-18
for desk phones 8-12
for LAN 3-14
for mobile clients and devices 21-80
for mobile endpoints 8-39
for music on hold 7-41
for security 4-31
for software-based endpoints 8-29
for Unified CM Assistant 18-32
for video 8-22, 13-78
for video endpoints 8-18
for WAN 3-33, 3-37
for wireless endpoints 8-36
for wireless LAN 3-74
quality of voice transmissions 7-39
queue, universal 22-7
queue depth 3-60
queuing of calls 22-2
queuing of voice traffic 3-18, 3-76
Quick Buffer Encoding (QBE) 9-29, 19-22

R

radio frequency (RF) 8-33
Rapid Spanning Tree Protocol (RSTP) 3-4, 3-7
rate of error 10-47
RBAC 27-4
RBOC 15-5
RCC 20-18, 20-62
RCP 4-11
RDNIS 19-7
real-time eventing model 20-55
Real Time Monitoring Tool (RTMT) 16-3
Real-Time Transport Protocol (RTP) 10-25
rebroadcast music on hold 7-22
recording
and silent monitoring 22-9, 23-4
calls 23-1
SPAN method 23-2
Redirected Dialed Number Information Service (RDNIS) 19-7
Redirector servlet 18-35
redundancy
call processing 9-14
cluster configurations 9-16
Extension Mobility 18-15
for messaging 19-17
for Mobile Voice Access 21-68
for presence servers 20-21
for remote sites 10-16
for Single Number Reach 21-58
for Unified CM Assistant 18-24
gateway support for 5-3, 5-9
IP Phone Services 18-5
load balancing 9-19
TFTP services 3-31
WebDialer 18-39
refresh tokens 16-57
Regional Bell Operating Company (RBOC) 15-5
regions
for call admission control 13-46, 13-47
maximum number 25-14
Remote Call Control (RCC) 20-18, 20-62
Remote Copy Protocol (RCP) 4-11

remote destination
caller ID matching 21-65
phone pickup 21-51, 21-64
profile 21-68
Remote Device 9-28
remote enterprise mobility 21-26
Remote Expert Solution 22-22
remote failover deployment model 10-54
Remote Monitoring (RMON) 27-9
remote site survivability 10-16
re-packetization of a stream 7-7
replication network 13-52
replication of database 9-9
Representational State Transfer (REST) 20-55
rerouting calling search space 21-68
resilience 9-1
Resource Reservation Protocol (RSVP) 3-34
REST 20-55
restrictions for
Extension Mobility 18-18
IP Phone Services 18-7
Unified CM Assistant 18-28
WebDialer 18-41
RF 8-33
RFC 2833 5-6, 7-7
rich media conferencing 11-1
Ring All Shared Lines 21-70
RIP 4-35
Rivest, Shamir, and Adelman (RSA) 4-16
RMON 27-9
roaming 3-69
Roaming Sensitive Settings 21-18
rogue
DHCP server 4-8
network extensions 4-8
role-based access control (RBAC) 27-4
roles in the network infrastructure 3-3
root guard 3-6
round-trip time (RTT) 10-46, 10-49
Routed Access Layer 3-7
routed ASA firewall 4-35
routers
access control list (ACL) 4-32
branch office 7-44
flash 7-44
roles and features 3-3
selective for E911 15-6
routes
filters 14-27
group devices 14-30
groups 14-28, 14-30
lists 14-29
patterns 14-22, 14-26
selection of 14-82
routing
calling line ID 14-28
calls 14-22, 21-80
digit manipulation 14-28
inbound calls 5-32
inter-VLAN 8-19, 8-30
least-cost 5-36
outbound calls 5-33
protocols 3-11
time-of-day (ToD) 14-91
Routing Information Protocol (RIP) 4-35
RSA 4-16
RSTP 3-4, 3-7
RSVP
WAN infrastructure 3-34
RTMT 16-3, 27-24
RTP 10-25
RTT 10-46, 10-49

S

SaaS 11-26
SAF
architecture 10-59

described 10-59
SAML 16-33, 16-35, 16-37, 20-7, 20-40, 21-105, 25-20
SAML bearer assertion grant flow 16-49
SAN 10-58
scalability of
IP Phone Services 18-6
Unified CM 9-1
scavenger class traffic 3-48
SCCP
DTMF signaling 7-8
gateway support for 5-3
music on hold (MoH) 7-23
phones 14-15
user input on phones 14-15
schema 16-1
scopes 16-57
SDK 16-4
search base for directories 16-13
Section 255 8-5
Section 508 8-5
Secure Mobility Client 8-38
secure mode for phones 18-14
Secure Real-time Transport Protocol (SRTP) 4-29
secure remote enterprise attachment 8-13, 8-19, 8-30, 8-38
security
access control list (ACL) 4-32
Cisco Unified Border Element 4-40
conferences 4-40
configuration example 4-43
data center 4-37
DHCP Snooping 4-8
DHCP starvation attack 4-10
directories 16-19
endpoints 4-25
Extension Mobility 18-13
firewalls 4-33, 4-44
gateways 4-38
infrastructure 4-4
in general 4-1, 4-2
intracluster communications 9-11
IPv6 addressing 4-5
layers 4-3
lobby phone example 4-43
MAC CAM flooding 4-7
media resources 4-38
PC port on the phone 4-26
phones 4-25
phone settings 4-28
physical access 4-4
policy 4-2
QoS 4-31
rogue network extensions 4-8
servers 4-41, 4-42
switch port 4-6
voice VLAN 4-26
VPN clients 4-30
web access 4-27
WebEx 20-67
Security Assertion Markup Language (SAML) 16-33, 16-35, 16-37, 20-40, 21-105, 25-20
Security Enhanced Linux (SELinux) 4-42
selecting the proper route 14-82
selective router 15-2, 15-6
SELinux 4-42
sending multicast voice messages 18-47
Sequenced Routing Update Protocol (SRTP) 3-53
servers
capacity planning 9-23, 25-14
clusters 9-5, 20-19
co-located 3-26
co-resident DHCP 3-27
co-resident MoH 7-31
CTI Manager 9-20
data center 3-12
farm 3-12
for DHCP 3-27
for media resources 7-1
for music on hold 7-31

for presence 20-18
multiple Unified CM servers 19-21
paging server 18-47
performance 9-23, 20-26
publisher 9-6, 10-45
redundancy 20-21
security 4-41, 4-42
standalone 3-27, 7-31
subscriber 9-6
synchronization of 20-19
TFTP 9-7, 9-20
Service Advertisement Framework (SAF)
architecture 10-59
described 10-59
service discovery 21-91
Service Inter-Working (SIW) 3-45, 10-15, 10-24
services
for IP phones 18-2
prefix 5-34
supplementary 5-5
within a cluster 9-5
service set identifier (SSID) 3-69, 3-72
servlet for
Redirector 18-35
WebDialer 18-34
Session Initiation Protocol (SIP)
annunciator 7-16
delayed offer 7-9
dial rules 14-20
early offer 7-9
for distributed call processing 10-25
for interoperability of Unified CM and Unified CME 9-40
gateways 5-11
gateway support for 5-6
music on hold (MoH) 7-26
phones 8-43, 14-16, 14-18
presence 20-14
trunks 6-3, 6-5, 6-6, 15-8
Type-A phones 14-16
Type-B phones 14-18
Session Management Edition (SME) 10-25, 10-26, 13-82
settings for IP phones 4-28
shadow location 13-55
shaping traffic 3-50
shared
line appearances 3-59, 15-22
line mode with Unified CM Assistant 18-21
locations 13-53
signaling encryption 3-58, 3-59
signal strength 5-32
silent monitoring and recording of calls 22-9, 23-4
SIMPLE 20-18
Simple Mail Transfer Protocol (SMTP) 19-28
Simple Network Management Protocol (SNMP) 15-10
Simple Object Access Protocol (SOAP) 20-19
simplified configuration for mobile client users 21-87
single-cluster deployment 20-26
single inbox 19-44
Single Number Reach (SNR) 21-46, 21-49
Single Sign On (SSO) 4-42, 11-54, 16-33, 16-35, 20-7, 20-66, 21-105
Single Sign-On (SSO) 20-4, 20-40, 25-20
single sign-on (SSO) 16-1
single site
deployment model 7-36, 7-43, 10-10, 22-12, 27-26
messaging model 19-4
Singlewire InformaCast 18-47
SIP
annunciator 7-16
delayed offer 7-9
dial rules 14-20
Early Offer 7-9
for distributed call processing 10-25
for interoperability of Unified CM and Unified CME 9-40
gateways 5-11
gateway support for 5-6

MTP requirements 7-11
music on hold (MoH) 7-26
phones 8-43, 14-16, 14-18
presence 20-14
route pattern 14-29
routing requests 14-48
trunks 6-3, 6-5, 6-6, 15-8
Type-A phones 14-16
Type-B phones 14-18
SIP for Instant Messaging and Presence Leveraging Extensions (SIMPLE) 20-18
site-based design 10-6
site survey for wireless network 8-33
SIW 3-45, 10-15, 10-24
sizing
Cisco Jabber clients 25-17
design and deployment considerations 25-1
factors to consider 25-9
methodology 25-2
tool 9-23, 25-10, 25-13
Unified CM servers 9-23
Skinny Client Control Protocol (SCCP)
DTMF signaling 7-8
gateway support for 5-3
music on hold (MoH) 7-23
phones 14-15
user input on phones 14-15
Smart Software Licensing 27-21
Smart Software Manager (SSM) 26-9, 27-21
SME 10-25, 10-26, 13-82, 25-12
SMTP 19-28
sn attribute 16-10
SNMP 15-10
snooping 4-8
SNR 21-46, 21-49
SOAP 20-19
SocialMiner 22-6, 22-7
softphone mode (audio on computer) 8-25
software
endpoints 8-22
media resource capacities 7-30
MTP resources 7-14
software as a service (SaaS) 11-26
Software Development Kit (SDK) 16-4
software eTokens 4-22
SPAN 23-2, 23-3
Spanning Tree Protocol (STP) 3-6
Spark 8-27, 8-37, 21-79, 21-108
Spark Calendar Service 21-38
Spark Call Service 21-41
Spark Hybrid Services 21-36
Spark Identity Service 21-36
Spark Room Series 8-17
speed dial presence 20-16
split tunneling 21-103
SRST 4-25, 7-44, 8-13, 8-19, 8-30, 8-36, 8-40, 8-42, 9-16, 10-15, 10-16, 10-19, 15-6
SRST Manager 10-21
SRSV 19-8
SRTP 3-53, 4-29
SSID 3-69, 3-72
SSM 26-9, 27-21
SSO 11-54, 16-1, 16-33, 16-35, 20-4, 20-7, 20-40, 20-66, 21-105, 25-20
standalone analog gateways 8-6
standalone server 3-27, 7-31
static ANI interface 15-15
static memory 25-4
stealth firewall 4-36
storage area networking (SAN) 10-58
STP 3-6
SUBSCRIBE calling search space 20-17
subscriber server 9-6
Sun ONE Directory Server 16-10, 16-16
supplementary services
design considerations 9-39
for H.323 endpoints 7-12
on gateways 5-5, 5-6

survey of wireless network 8-33
Survivable Remote Site Telephony (SRST) 4-25, 7-44, 8-13, 8-19, 8-30, 8-36, 8-40, 8-42, 9-16, 10-15, 10-16, 10-19, 15-6
Survivable Remote Site Telephony (SRST) Manager 10-21
Survivable Remote Site Voicemail (SRSV) 19-8
Switched Port Analyzer (SPAN) 23-2, 23-3
switches
  port security 4-6
  roles and features 3-3
switch port discovery 15-11
SX Series video endpoints 8-17
synchronization of
  directories 16-10
  presence servers 20-19
  Unified CM database 16-31
system memory 25-4

T

Tail End Hop Off (TEHO) 14-71
TAPI 9-20
TEHO 14-71
Telecommunications Act 8-5
telephone record and playback (TRaP) 19-4
telephone user interface (TUI) 19-4
Telephony Service Provider (TSP) Audio 11-57
TelePresence
  call admission control 13-59
  call routing 14-53
  dial plan 14-53
  endpoints 4-28, 8-16, 8-17, 8-18
  interoperability 8-20, 13-78
  Quality of Service (QoS) 8-22, 13-78
TelePresence ISDN Link 5-3
TelePresence Management Suite (TMS) 11-52, 16-33, 27-18
TelePresence Management Suite Extension Booking API (TMSBA) 27-18
TelePresence Management Suite Extension for IBM Lotus Notes (TMSXN) 27-18
TelePresence Management Suite Extension for Microsoft Exchange (TMSXE) 11-52, 27-18
TelePresence Management Suite Provisioning Extension (TMSPE) 11-52, 27-20
termination of calls 7-4
test calls for 911 15-21
Tested Reference Configuration (TRC) 9-4, 10-55
Text Conference Manager 20-41
TFTP 3-25, 3-28, 4-24, 9-5, 9-20
third-party
  IP phones 8-43
  SIP phones 8-43
third-party CA certificates 4-30
Third-Party Open API 20-55
third-party XMPP clients 20-69
third-party XMPP clients and applications 25-21
time-of-day (ToD) routing 14-91
timer control mobile voicemail avoidance 21-55
timers for call signaling 5-14
time synchronization 3-33
TLS 4-20, 4-29
TMS 11-52, 16-33, 27-18
TMSBA 27-18
TMSPE 11-52, 27-20
TMSXE 11-52, 27-18
TMSXN 27-18
ToD 14-91
tokenless 4-22
tokens 16-51, 16-56
toll fraud mitigation 5-24
topology for call admission control 13-73
TPC 3-69
tracking domain 15-25, 15-26
traffic
  bearer traffic 3-53
  call control 3-57, 3-61
  classification 3-4, 3-16, 3-75
  conferencing and collaboration 25-8
  contact centers 25-7

engineering 25-5, 25-6
planning for WebEx 11-33
prioritization 3-47
provisioning for 3-53
PSTN traffic patterns 25-39
queuing 3-18, 3-76
shaping 3-50
video bearer traffic 3-56
video calls 25-7
voice bearer traffic 3-53, 25-6
voice calls 25-6
transcoding
Cisco Unity 19-33
described 7-5
resources 7-6
transformations
caller ID 21-72
of calling and called numbers 14-58
translation of digits 14-24
translation patterns 14-24
Transmit Power Control (TPC) 3-69
transparent ASA firewall 4-36
Transport Layer Security (TLS) 4-20, 4-29
TRaP 19-4
TRC 9-4, 10-55
Trivial File Transfer Protocol (TFTP) 3-25, 3-28, 4-24, 9-5, 9-20
troubleshooting for clustering over the WAN 10-47
TRP 3-18, 7-15
trunks
architecture 6-2
comparing H.323 and SIP 6-3
described 6-1
features supported 6-3
SIP 6-5, 6-6, 7-16, 15-8
utilization of 27-10
Trusted Relay Point (TRP) 3-18, 7-15
TSP Audio 11-57
TUI 19-4
two-stage dialing 21-63, 21-65
Type-A phones 14-16
Type-B phones 14-18

U

UCS
high availability 9-21
QoS 3-20
virtualized servers 10-55
UDLD 3-6
UDP 3-48, 10-25
UDS 8-26, 16-6, 16-32, 20-8, 21-93
UDS proxy for LDAP 16-32
UN 5-6
unassigned DNs 14-68
unicast call flow 7-25, 7-28
unicast music on hold 7-17, 7-23, 7-40
UniDirectional Link Detection (UDLD) 3-6
Unified Analysis Manager 27-24
Unified Border Element 4-40
Unified CCE 22-3
Unified CCMP 22-8
Unified CCX 22-6
Unified CM
capacity planning 25-13
database synchronization 16-31
groups 10-48, 10-54
hardened platform 4-21
mixed mode 4-22
presence 20-14
sizing tool 9-23
Unified CM Assistant 18-19, 25-27
Unified CME
capacity planning 9-26, 25-49
design considerations 9-28
distributed call processing 10-25
interoperability with Unified CM 9-36
Unified CM Express (Unified CME)

capacity planning 9-26, 25-49
design considerations 9-28
distributed call processing 10-25
interoperability with Unified CM 9-36
Unified Communications Management Suite 27-1
Unified Communications Manager Assistant (Unified CM Assistant) 18-19
Unified Communications Manager Real-Time Monitoring Tool (RTMT) 27-24
Unified Communications System
applications and services layer 17-1
call routing layer 12-1
introduction 1-1
operations and serviceability layer 24-1
Unified Computing System (UCS)
high availability 9-21
QoS 3-20
virtualized servers 10-55
Unified Contact Center 22-1
Unified Contact Center Enterprise (Unified CCE) 22-3
Unified Contact Center Express (Unified CCX) 22-6
Unified Contact Center Management Portal (Unified CCMP) 22-8
Unified Customer Voice Portal (Unified CVP) 22-4
Unified CVP 22-4
Unified EIM 22-9
Unified E-Mail Interaction Manager (Unified EIM) 22-9
Unified IC 22-9
Unified Intelligence Center (Unified IC) 22-9
Unified MeetingPlace 25-44, 25-45
unified messaging (*see also* messaging) 19-1
Unified Mobility 21-1, 21-47, 21-68, 21-107, 25-21, 25-52
Unified PM 27-13
Unified Provisioning Manager (Unified PM) 27-13
Unified Reporting 27-25
Unified Service Monitor (Unified SM) 27-8
Unified SM 27-8
Unified Survivable Remote Site Telephony (SRST) 10-16
Unified Web Interaction Manager (Unified WIM) 22-9
Unified WIM 22-9
uninterrupted power supplies (UPS) 3-12
Unity 19-1, 19-6, 19-19
Unity Connection 19-6, 19-17
Unity Express 19-22
Unity Telephony Integration Manager (UTIM) 19-40, 19-42
Universal Line Template 16-17
Universal Queue 22-7
Unsolicited Notify 7-8
Unsolicited SIP Notify (UN) 5-6
UP 3-75
UplinkFast 3-6
UPS 3-12
URI dialing 14-23, 14-49, 14-50, 14-52
URLs for WebDialer 18-38
user authentication 16-45
user control mobile voicemail avoidance 21-56
User Datagram Protocol (UDP) 3-48, 10-25
User Data Service (UDS) 8-26, 16-6, 16-32, 20-8, 21-93
user hold 7-19
UserID 16-10
user priority (UP) 3-75
users
application users 16-7
directory search base 16-13
end users 16-7
input on phones 14-15, 14-16, 14-18
UTIM 19-40, 19-42

V

V3PN 10-15, 10-24
VAD 25-40
VAF 3-49
variable length on-net dial plan 21-23
VATS 3-51
VCS
dial plan 14-53
directory integration 16-33

integration with Unified CM 14-74
TelePresence Management Suite (TMS) 27-18
VDS 23-3
video
bandwidth utilization 13-66
bearer traffic 3-56
call admission control 13-66, 13-78
customer care 22-22
endpoints 8-14, 15-20
gateways 5-11
interoperability 7-6, 8-20, 13-78
migration to Unified CM 26-11
over wireless LAN (WLAN) 21-78
Quality of Service (QoS) 8-22, 13-78
traffic characteristics 25-7
traffic classification 3-17
VLAN 4-5
Video Communication Server (VCS)
dial plan 14-53
directory integration 16-33
integration with Unified CM 14-74
TelePresence Management Suite (TMS) 27-18
videoconferencing 25-45
ViewMail for Outlook (VMO) 19-4
virtualization
of call processing 9-3
of Cisco Unity Connection 19-31
Virtualization Experience Media Engine (VXME) 8-42
virtualized servers 10-55, 26-9
Virtualized Voice Browser (VVB) 22-12
virtual LAN (VLAN) 3-4, 3-69
virtual machine 26-9
Virtual Private Network (VPN) 10-15, 10-24
Virtual Router Redundancy Protocol (VRRP) 3-9
virtual software switches 3-20
visiting cluster 18-14
VLAN
access control list (ACL) 4-32
number of devices per VLAN 3-4
separate VLANs for voice and data 3-69
video 4-5
voice 4-5, 4-26
VMO 19-4
VMware 3-20, 10-55
voice
bandwidth requirements 3-49
bearer traffic 3-53
gateways 5-1, 8-5
port integration 19-40, 19-42
termination 7-4
traffic 25-6
VLAN 4-5, 4-26
voice activity detection (VAD) 25-40
Voice-Adaptive Fragmentation (VAF) 3-49
Voice-Adaptive Traffic Shaping (VATS) 3-51
Voice and Video Enabled IPSec VPN (V3PN) 10-15, 10-24
voicemail
avoidance 21-55
Cisco Unity 19-1
Cisco Unity Express 19-22, 19-28
for local failover 10-53
mobile users 21-55
networking 19-28
single inbox 19-44
third-party systems 19-47
unified messaging 19-1
with Single Number Reach 21-55
voice messaging 19-1, 25-42
voice over IP (VoIP) 3-53
voice over the PSTN (VoPSTN) 10-22
Voice Profile for Internet Mail (VPIM) 19-28
voice quality 7-39
voice quality monitoring 27-8, 27-10
voice rtp send-recv command 15-18
voice traffic 25-6
VoiceXML (VXML) 21-60, 21-61
VoIP 3-53
VoPSTN 10-22

VPIM 19-28
VPN 4-30, 10-15, 10-24
VPN-less access 10-36
VPN-less secure remote connectivity 21-30
VRRP 3-9
vSphere 3-20
vSphere Distributed Switch (VDS) 23-3
VVB 22-12
VXI 25-13
VXME 8-42
VXML 21-60, 21-61

W

WAN
  aggregation router 3-3
  infrastructure 3-33
watcher lists 20-59
web access from IP phone 4-27
WebDialer 18-34, 25-27
WebEx 8-26, 11-26, 20-10, 21-79
WebEx Collaboration Cloud 11-26
WebEx Connect 25-20
WebEx Meeting Center Video Conferencing 11-34
WebEx Meetings 8-38, 21-108
WebEx Meetings Server 11-41
WebEx Messenger 20-64
weighted fair queuing 3-47
weights for call admission control 13-41
WEP 8-34
what's new for this release
  call admission control 13-1
  call processing 9-2
  call recording and monitoring 23-1
  Cisco Unified Contact Center 22-2
  deployment models 10-1
  dial plan 14-2
  endpoints 8-2
  gateways 5-1
  LDAP directory integration 16-2
  mobility applications 21-3
  network infrastructure 3-4
  network management 27-2
  presence 20-2
  rich media conferencing 11-3
  security 4-1
  sizing considerations 25-2
  system migration 26-2
  Unified CM applications 18-2
white list 20-68
Wi-Fi Multimedia (WMM) 3-76
Wi-Fi Multimedia Traffic Specification (WMM TSPEC) 3-77
wildcard route pattern 14-26, 14-27
Windows Internet Naming Service (WINS) 3-27
wink start 8-6
WINS 3-27
Wired Equivalent Privacy (WEP) 8-34
wireless
  access points 3-63
  endpoints 3-65, 8-33
  IP Phone 7921G 8-33
  IP Phone 7925G 8-33
  IP Phone 7925G-EX 8-33
  IP Phone 7926G 8-33
  IP phones 8-33
  LAN 3-61
  LAN controller (WLC) 3-64, 3-73
wireless LAN (WLAN) 3-61, 8-38
WLAN infrastructure 3-61, 8-38
WLC 3-64, 3-73
WMM 3-76
WMM TSPEC 3-77

X

XCP Text Conference Manager 20-41
XMPP clients 20-69, 25-21